arXiv:2601.12962v2 [cs.CY] 30 Aug 2026

ACE-Align: Attribute Causal Effect Alignment for Cultural Values under Varying Persona Granularities

Jiatang Luo[1,2,3], Bingbing Xu[2,3]*, Rongxin Chen[2,3], Xiaoyan Zhao[4,5], Yang Zhang[4], Liang Pang[2,3], Zhiyong Huang[4], Huawei Shen[2,3],
[1]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, China
[2]State Key Laboratory of AI Safety, Institute of Computing Technology, CAS
[3]University of Chinese Academy of Sciences
[4]National University of Singapore
[5]The Chinese University of Hong Kong
luojiatang23@mails.ucas.ac.cn,{xubingbing,chenrongxin24s,pangliang}@ict.ac.cn

Abstract

Ensuring that large language models (LLMs) reflect diverse cultural values is important for globally deployed NLP systems. However, existing approaches often treat cultural groups as homogeneous and overlook within-group heterogeneity arising from intersecting demographic attributes, leading to unstable behavior under varying persona granularity. To address this gap, we propose **ACE-Align** (**A**ttribute **C**ausal **E**ffect Alignment), a causally inspired framework based on controlled persona edits that aligns how specific demographic attributes shift different cultural values, rather than treating each culture as a homogeneous group. We evaluate ACE-Align across 14 countries spanning five continents, with personas specified by subsets of four attributes (gender, education, residence, and marital status) and granularity instantiated by the number of specified attributes. Across all persona granularities, ACE-Align consistently outperforms baselines. Moreover, in within-survey comparisons, it reduces the average Global North–South alignment gap from 3.40 to 1.11 points on WVS and from 2.53 to 0.85 points on ISSP. Code and dataset are released at this link.

1 Introduction

As large language models (LLMs) become globally deployed, aligning their outputs with diverse cultural values has become an important challenge for socially responsible NLP systems (Cao et al., 2023). Despite substantial progress in instruction tuning and preference optimization, a persistent alignment deficit remains: a model may appear aligned with a broad cultural or national group, yet still misrepresent sub-populations within that group that are defined by intersecting demographic attributes (Santurkar et al., 2023). This deficit is especially consequential in regions such as Africa and Southeast Asia, where uneven digital access and scarce high-quality text contribute to weaker representation in the corpora used to train LLMs, increasing the risk of cultural erasure and stereotyping (Decoupes et al., 2025).

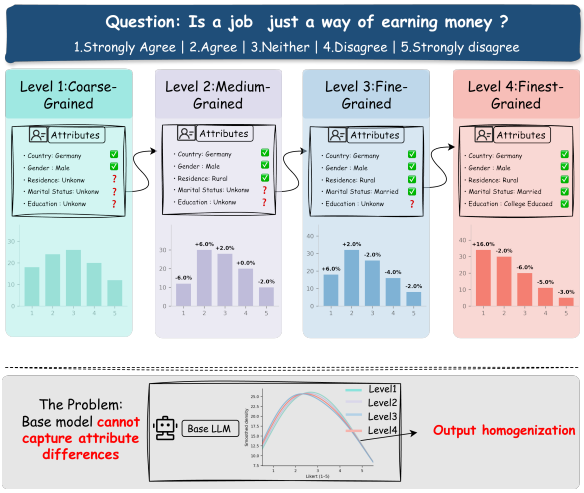


Figure 1: Within a single culture, personas at different granularities exhibit heterogeneous values, which current LLMs fail to consistently capture.

This challenge becomes more pronounced when persona conditioning varies in granularity. As illustrated in Figure 1, personas may describe broad sub-populations using country and a single demographic attribute, or progressively narrower sub-populations by combining multiple intersecting attributes (Chen et al., 2026). A model may behave plausibly for a coarse-grained persona (e.g., *German man*), yet degrade or become inconsistent when conditioned on a finer-grained persona that adds residence, education, and marital status (e.g., *a rural, college-educated, married German man*). Consistent with this challenge, Table 1 shows that all evaluated methods obtain lower average alignment scores at $G{=}4$ than at $G{=}1$. This suggests that cultural alignment methods should remain reliable across persona granularities, rather than being optimized for a single granularity level.

* Corresponding author.

Existing alignment strategies often fit a direct mapping from persona descriptions to target responses, without explicitly constraining how responses should shift when persona attributes are combined (Xu et al., 2025; AlKhamissi et al., 2024). This can lead to unstable attribute interactions: models may exhibit *stereotyping* (over-amplifying demographic influences) (Anthis et al., 2025), *erasure* (understating or homogenizing demographic differences) (Santurkar et al., 2023), or *flipped signs* (reversing the direction of demographic influence) (Talat et al., 2022).

To address these failures, we argue that alignment across persona granularities requires explicitly modeling how demographic attributes are associated with shifts in cultural values. Inspired by counterfactual reasoning in causal inference, we operationalize this with controlled persona edits: one attribute is toggled while the country, question, and other specified attributes are held fixed. These contrasts provide survey-grounded estimates of how each attribute is associated with response shifts, but they should not be interpreted as randomized real-world intervention effects.

In this paper, we instantiate this perspective with **ACE-Align** (**A**ttribute **C**ausal **E**ffect Alignment), which trains LLMs to match model-side controlled persona-edit effects with survey-grounded attribute effects. Since effect matching only constrains relative response shifts, it can admit multiple absolute prediction distributions with the same attribute effect. ACE-Align therefore adds a lightweight anchoring loss that grounds absolute predictions while preserving the learned response shifts. This encourages the model to match both the direction and magnitude of attribute influence while remaining calibrated to survey responses.

Our contributions are summarized as follows:

- **Investigating Persona Granularity.** We formalize the challenge of cultural alignment across varying persona granularities and reveal the performance instability of LLMs under intersecting demographic attributes.
- **ACE-Align Framework.** We propose **ACE-Align**, which matches controlled persona-edit effects to survey-grounded attribute effects while anchoring absolute predictions, improving consistency across persona granularities.
- **Global Equity and Bias Diagnosis.** Following the Brandt-line North–South division (Brandt et al., 1980), we compare comparatively industrialized economies (Global North) with developing and emerging economies (Global South). Within WVS and ISSP separately, ACE-Align narrows the resulting alignment gap and reduces attribute-effect errors such as stereotyping and cultural erasure.

2 Related Work

2.1 Cultural Alignment

Recent work evaluates and improves the cultural alignment of LLMs using survey-style benchmarks and cross-cultural opinion distributions (Cao et al., 2023; Santurkar et al., 2023; Wang et al., 2025; Sukiennik et al., 2025; Kovač et al., 2023; Tao et al., 2024). Existing approaches span prompting and inference-time strategies (AlKhamissi et al., 2024) and alignment training via supervised fine-tuning on survey data (Li et al., 2024; Masoud et al., 2025; Du et al., 2025). These methods often target culture at a fixed resolution, which can mask within culture heterogeneity when persona attributes are specified at different granularity levels (Decoupes et al., 2025). Our work focuses on consistency across persona granularities and treats demographic attributes as compositional factors that drive heterogeneous cultural topic preferences.

2.2 Causal Perspectives for Debiasing

Causal formulations have been used to analyze and mitigate bias by comparing interventions and counterfactual contrasts rather than relying only on correlations (Chaudhary et al., 2025). Prior work has applied this perspective to dataset bias detection via causal invariance and active selection (Sun et al., 2024), reasoning debiasing through counterfactual chains (Wu et al., 2024), and annotation artifact mitigation through causal mediation analysis and targeted unlearning (Lim et al., 2024). Other studies localize where social bias is mediated inside language models (Vig et al., 2020), or use front-door adjustment for prompting-time interventions when model parameters are inaccessible (Zhang et al., 2025). Our work differs by applying controlled persona edits to cultural alignment under persona conditioning, align-

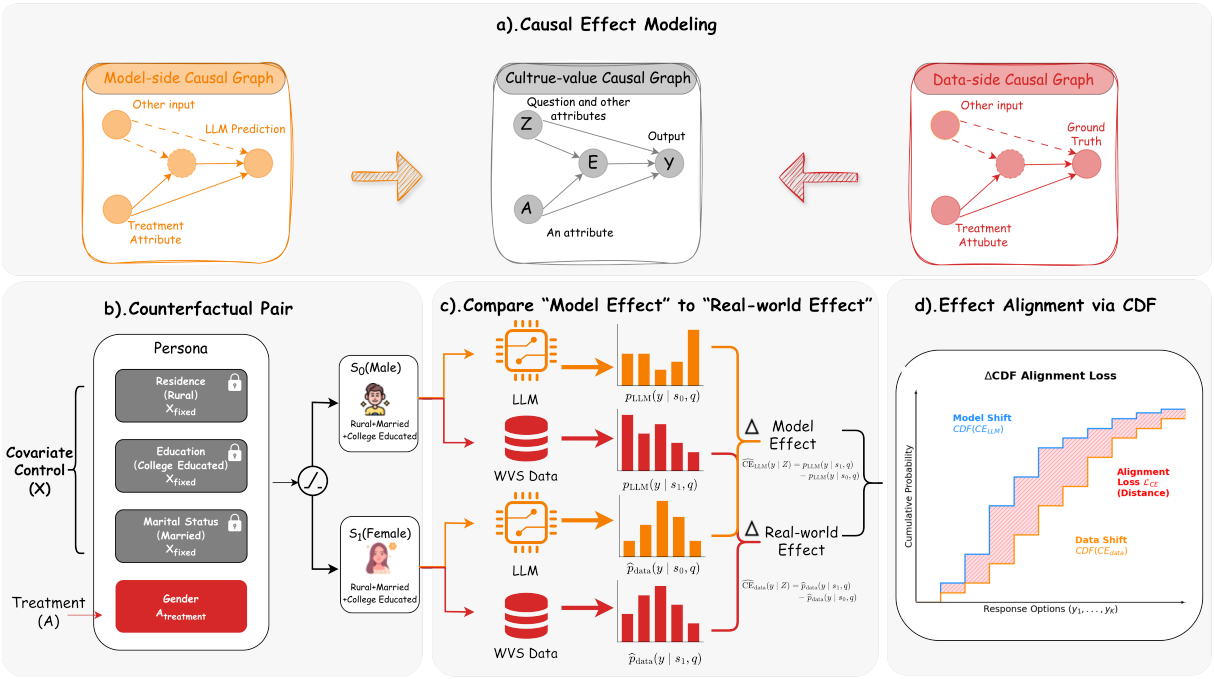


Figure 2: **The ACE-Align framework.** We optimize cultural alignment at the finest granularity ($G = 4$) to keep the remaining persona covariates fixed when constructing attribute-effect contrasts. By aligning the model's controlled response shifts ($\mathrm{CE}_{\mathrm{LLM}}$) with survey-grounded attribute effects ($\mathrm{CE}_{\mathrm{data}}$) via ΔCDF distance, the model learns response-shift patterns that maintain consistency across persona granularities.

ing attribute-level response shifts to improve consistency across persona granularities.

3 Method

3.1 Task Formulation

We investigate cultural alignment of model responses to culture topics under persona conditioning at varying granularity. Here, a persona denotes a subgroup profile defined by a set of attributes, rather than a specific individual identity. We instantiate persona granularity by the number of specified attributes in the profile.

Let $\mathcal{A} = \{a_1, \ldots, a_M\}$ denote the universe of available attributes. A persona g is an instantiated subset of $\mathcal{A}$, where selected attributes are assigned specific values. The persona granularity is defined as the number of specified attributes, denoted by the scalar $G = |g|$.

Given a value-laden question q, the response options form an ordered discrete set $\mathcal{Y} = \{1, \ldots, K\}$. The model θ outputs a categorical distribution over these options:

$$p_\theta(y \mid g, q), \quad y \in \mathcal{Y}, \tag{1}$$

where θ represents the model parameters.

Obtaining option probabilities from the LLM. To instantiate the categorical distribution over the discrete option set $\mathcal{Y}$, we compute the likelihood of each option under the LLM and renormalize over $\mathcal{Y}$. Specifically, for each option $y \in \mathcal{Y}$, let $\mathbf{t}(y) = (t_1, \ldots, t_{m_y})$ be the token sequence of its canonical verbalization under the model tokenizer. Given the formatted prompt(g, q), we define the option log-likelihood as

$$\ell_\theta(y \mid g, q) = \sum_{j=1}^{m_y} \log p_\theta(t_j \mid \text{prompt}(g, q), t_{<j}), \tag{2}$$

and obtain a categorical distribution by renormalizing within $\mathcal{Y}$:

$$p_\theta(y \mid g, q) = \frac{\exp(\ell_\theta(y \mid g, q))}{\sum_{y' \in \mathcal{Y}} \exp(\ell_\theta(y' \mid g, q))}. \tag{3}$$

3.2 ACE-Align Framework

To capture how demographic attributes influence different cultural topics, we propose to align attribute causal effects. By learning survey-grounded response shifts associated with each attribute, the model can compositionally generalize to novel attribute combinations. As illustrated in Figure 2, we introduce **ACE-Align** (**A**ttribute

Causal **E**ffect Alignment), a framework that explicitly estimates and aligns these effects during training. In training, we instantiate contexts using the finest granularity level ($G = 4$), so that when toggling a treatment attribute A, all remaining demographic attributes are held fixed in X. This provides the strongest available covariate control in our persona design and reduces variation from omitted specified attributes, yielding more reliable shifts associated with the toggled attribute.

3.2.1 Causal Effect Modeling

Causal Graph. For a culture topic question q with ordinal options $\mathcal{Y}$, we treat the selected option as a discrete outcome variable $Y \in \mathcal{Y}$. We choose one attribute $A \in \mathcal{A}$ as the treatment and denote the remaining persona specification as X. We define the context as $Z = (X, q)$, where X contains all other specified attributes and q specifies the culture topic question. Figure 2a shows our directed acyclic graph (DAG) (Pearl, 2009), where A and Z directly influence Y. To account for complex and often unobservable dependencies between demographics and cultural contexts, we include a latent factor E that represents interactions between A and Z (Zhao et al., 2026). We operationalize the attribute effect by toggling only A while holding Z fixed.

Identification Assumption. For each treated attribute A, we view controlled persona editing as an operational intervention on the persona specification: $\mathrm{do}(A = a)$ sets the value of one persona attribute while keeping $Z = (X, q)$ fixed. In the survey data, we approximate this controlled-edit distribution by the corresponding conditional distribution:

$$p_a^{\mathrm{do}}(y \mid Z) \approx p(Y = y \mid A = a, Z). \quad (4)$$

Importantly, our causal interpretation is defined in the persona-conditioning space: on the model side, changing A while holding Z fixed is a controlled input intervention, whereas on the survey side, because demographic attributes are not randomized, we use matched conditional response distributions as survey-grounded approximations of the corresponding controlled-edit effects. This approximation supports controlled persona-edit contrasts, but should not be interpreted as identifying randomized real-world causal effects.

Binary treatment coding. For each treatment attribute, we define a binary variable $A \in \{0, 1\}$ and estimate attribute effects via controlled persona edits. We consistently define the causal effect direction as $p(Y \mid A{=}1, Z) - p(Y \mid A{=}0, Z)$. The concrete attribute-value coding is reported in Appendix B.

3.2.2 Attribute Causal Effect

For a target attribute A under a fixed context $Z = (X, q)$, let $r \in \{m, d\}$ index the model and survey data, with $p_m = p_\theta$ and $p_d = p_{\text{data}}$. For compactness, write $p_r^a(y \mid Z) = p_r(Y = y \mid \mathrm{do}(A = a), Z)$. For each answer option $y \in \mathcal{Y}$ and $r \in \{m, d\}$, the option-level causal effect is:

$$\mathrm{CE}_r(y \mid A, Z) = p_r^1(y \mid Z) - p_r^0(y \mid Z). \quad (5)$$

Under Eq. (4), we operationalize this effect with two toggled persona specifications $s_1 = (A = 1, X)$ and $s_0 = (A = 0, X)$:

$$\widehat{\mathrm{CE}}_r(y \mid A, Z) = \widehat{p}_r^1(y \mid Z) - \widehat{p}_r^0(y \mid Z). \quad (6)$$

Here $\widehat{p}_r^a(y \mid Z) = \widehat{p}_r(y \mid s_a, q)$. For the model, $\widehat{p}_m$ is obtained from a forward pass; for survey data, $\widehat{p}_d$ is the empirical response distribution of matched respondents.

3.2.3 Effect Alignment Objective

Aligning Attribute Causal Effects. The core of our framework is to ensure the model mimics how human responses *shift* when a specific attribute A is toggled. This is achieved via three steps:

- **Step 1: Vector Representation.** For each (A, Z) and $r \in \{m, d\}$, we stack the option-level effects from Eq. (6) into a vector over $k = 1, \ldots, K$:

$$\widehat{c}_{r,k}(A, Z) = \widehat{\mathrm{CE}}_r(k \mid A, Z). \quad (7)$$

- **Step 2: Cumulative Transformation.** Since the response options $\mathcal{Y}$ are ordinal, we cumulatively sum the option-level effects to capture distributional shifts along the ordinal scale:

$$\Delta F_r(k \mid A, Z) = \sum_{i=1}^{k} \widehat{c}_{r,i}(A, Z). \quad (8)$$

- **Step 3: Measuring Alignment.** We measure the distance between model and data shifts across cumulative levels:

$$d_{\mathrm{CDF}}(A, Z) = \frac{1}{K} \sum_{k=1}^{K} \big| \Delta F_m(k \mid A, Z) - \Delta F_d(k \mid A, Z) \big|. \quad (9)$$

The total attribute-effect matching loss $\mathcal{L}_{\text{CE}}$ is defined as:

$$\mathcal{L}_{\text{CE}} = \frac{1}{|\mathcal{D}|} \sum_{(A,Z)\in\mathcal{D}} d_{\text{CDF}}(A, Z). \tag{10}$$

Here $\mathcal{D}$ denotes the set of valid (A, Z) pairs, where A is a demographic attribute and $Z = (X, q)$ is a context. We include (A, Z) only when both toggled personas $s_1 = (A{=}1, X)$ and $s_0 = (A{=}0, X)$ are supported by sufficient survey respondents for question q. Concretely, we discard a context if either subgroup has fewer than 10 matched respondents. In training, we construct $Z = (X, q)$ using the finest-grained personas ($G = 4$), so that when toggling A all remaining demographic attributes are held fixed in X, matching the controlled edit construction described above.

Anchoring the Model Response. The attribute-effect matching loss $\mathcal{L}_{\text{CE}}$ constrains the relative shift between a toggled persona pair under a fixed context, but is under-determined when optimized alone, as multiple absolute distributions can realize the same shift. To provide absolute reference points for model predictions, we introduce a per-persona anchoring loss:

$$\ell_{\text{anchor}}(s, q) = -\log p_\theta(y^*_{s,q} \mid s, q), \tag{11}$$

where $y^*_{s,q}$ denotes the empirical mode of the survey responses under persona s and question q. For each $(A, Z) \in \mathcal{D}$, this loss is applied to both endpoints $s_1 = (A{=}1, X)$ and $s_0 = (A{=}0, X)$ of the controlled persona edit.

Overall Objective. The final training objective integrates absolute response grounding with attribute-effect matching:

$$\mathcal{L} = \alpha\, \mathcal{L}_{\text{anchor}} + \beta\, \mathcal{L}_{\text{CE}}, \tag{12}$$

where α and β balance absolute response grounding and attribute-effect matching. We define $\mathcal{L}_{\text{anchor}}$ by aggregating $\ell_{\text{anchor}}(s, q)$ over both endpoints s_1 and s_0 for each $(A, Z) \in \mathcal{D}$.

4 Experiments

In this section, we report comprehensive experimental results on cultural alignment under persona conditioning at varying persona granularities. Our analysis is guided by the following three research questions **(RQs)**:

- **RQ1: Does ACE-Align consistently improve cultural alignment across varying persona granularities?**
- **RQ2: How robust is ACE-Align under low-resource conditions?**
- **RQ3: What mechanism drives the effectiveness of ACE-Align?**

4.1 Settings

Datasets. Following prior work (Cao et al., 2023; Masoud et al., 2025; Zhao et al., 2024; Sukiennik et al., 2025), we assess LLM cultural alignment by subjecting models to standardized sociological surveys under diverse demographic contexts. We use the *World Values Survey* (WVS) Wave 7 as training data (Haerpfer et al., 2022). For evaluation, we test on ISSP (Wang et al., 2025) for countries in America, Asia, Europe, and Oceania. For the three African countries, we instead evaluate on WVS using an 80/20 split, due to the lack of comparable ISSP coverage. To verify that this coverage-driven protocol difference does not drive the main conclusion, Appendix H reports WVS 80/20 held-out results for the non-African countries as a supplementary robustness check. We conduct experiments on 14 countries across five continents.[1] Appendix A reports topic definitions, 13 WVS and 10 ISSP topics, and per-country sample sizes. We consider four binary persona attributes, namely *Gender*, *Education*, *Residence*, and *Marital Status*, as these attributes are known to systematically relate to cultural values and attitudes in cross-national survey research (Kenny and Patel, 2017; Lomazzi and Seddig, 2020; Serrano-Montilla et al., 2020). Country is treated as part of the context X.

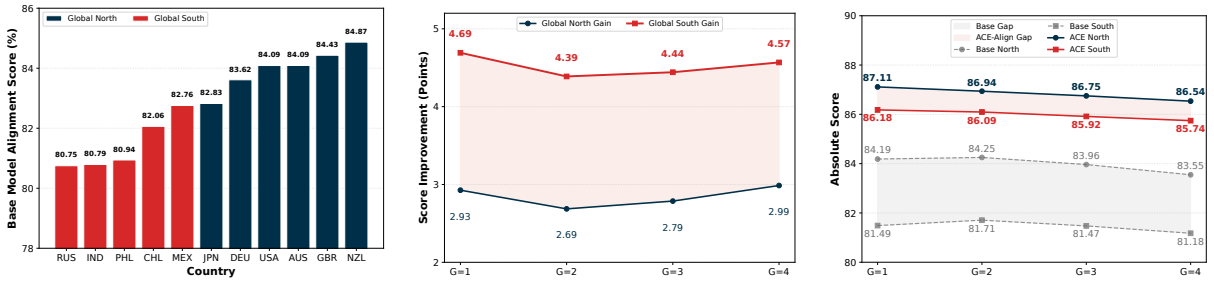
Baselines. Table 1 compares ACE-Align against external baselines and internal ablations, all instantiated on LLaMA3.1-8B-Instruct (Dubey et al., 2024). Additional results on Qwen3-8B (Yang et al., 2025) and Llama-3.1-70B-Instruct are reported in Appendix G. **(1) Base model:** LLaMA3.1-8B-Instruct is evaluated using the same persona-conditioned prompting protocol, without any additional alignment training. **(2) Prompt-based Method:** *Anthropological Prompting* augments prompting with an explicit anthropological reasoning framework before selecting an

[1] Country identifiers follow ISO-3166-1 alpha-3. The specific abbreviations and continents are listed in Appendix A.

	America			Europe			Asia			Africa			Oceania		Avg
	CHL	MEX	USA	DEU	GBR	RUS	IND	JPN	PHL	EGY	ETH	NGA	AUS	NZL	
$G = 1$															
Base model	82.16	82.75	84.04	83.74	84.71	80.46	81.37	82.89	80.70	72.87	74.74	77.35	84.44	85.30	81.25
AdPrompt	83.20	83.37	84.37	83.82	84.47	81.47	82.36	83.03	81.53	73.05	74.93	78.36	84.15	84.86	81.64 (+0.39)
DPO	81.70	82.75	83.16	83.69	84.25	81.01	81.56	82.90	81.42	74.35	76.53	79.20	84.03	85.29	81.56 (+0.31)
PPO	81.27	81.22	82.61	82.78	83.32	81.20	80.57	82.04	81.55	72.10	74.03	76.11	82.94	84.19	80.42 (-0.83)
GRPO	82.35	80.38	84.92	83.92	85.92	80.56	84.73	82.90	83.28	80.88	**85.49**	**86.12**	86.10	86.31	83.85 (+2.60)
SimLLM	83.55	84.63	86.21	85.49	86.43	83.12	83.20	84.52	83.25	80.11	83.48	85.19	86.48	**87.17**	84.49 (+3.24)
CultureSPA	81.72	83.68	83.97	82.44	83.20	81.67	83.23	82.25	82.04	73.48	77.50	81.13	82.95	83.26	81.61 (+0.36)
Anchor-only	84.18	79.58	81.63	81.91	80.03	79.82	80.60	81.77	82.49	79.58	73.56	74.62	82.12	82.30	80.30 (-0.95)
Causal-only	83.65	82.86	86.21	85.53	84.84	82.28	86.37	81.69	84.50	72.81	76.68	79.56	84.65	85.92	82.68 (+1.43)
ACE-Align	**85.17**	**86.39**	**87.75**	**87.54**	**87.23**	**84.53**	**87.57**	**85.19**	**87.24**	**85.00**	79.33	84.25	**87.98**	86.99	**85.87** (+4.62)
$G = 2$															
Base model	82.24	82.94	84.31	83.81	84.69	81.05	81.19	83.07	81.11	72.92	74.12	76.95	84.39	85.24	81.29
AdPrompt	83.10	83.22	84.24	83.80	84.38	81.60	82.00	83.08	81.52	73.24	74.46	78.00	84.08	84.76	81.53 (+0.24)
DPO	81.76	82.81	83.59	83.81	84.32	81.60	81.38	83.03	81.82	74.69	75.82	78.45	83.94	85.26	81.59 (+0.30)
PPO	81.20	80.89	82.37	82.62	83.11	81.35	80.19	81.66	81.60	72.20	73.69	76.13	82.60	83.80	80.24 (-1.05)
GRPO	81.90	79.54	84.83	83.65	85.78	80.45	84.21	82.83	83.23	81.16	**85.15**	**85.84**	85.90	86.09	83.61 (+2.32)
SimLLM	83.51	84.59	86.23	85.44	86.28	83.45	82.90	84.61	83.57	79.93	82.93	84.70	86.25	**86.95**	84.38 (+3.09)
CultureSPA	81.58	83.26	83.64	82.21	82.87	81.62	82.82	82.07	82.10	73.58	77.30	80.68	82.70	83.00	81.39 (+0.10)
Anchor-only	83.28	79.19	81.89	81.73	80.24	79.92	79.87	81.24	82.26	79.52	72.85	74.15	81.91	82.16	80.02 (-1.27)
Causal-only	83.78	82.79	86.15	85.60	85.00	82.59	86.35	81.75	84.98	73.08	76.27	79.31	84.77	86.15	82.76 (+1.47)
ACE-Align	**85.03**	**86.05**	**87.60**	**87.47**	**87.13**	**84.79**	**87.19**	**85.02**	**87.41**	**85.19**	79.19	84.74	**87.68**	86.73	**85.80** (+4.51)
$G = 3$															
Base model	82.03	82.70	84.18	83.59	84.40	80.88	80.67	82.80	81.09	72.84	73.54	76.39	84.02	84.79	80.99
AdPrompt	82.82	82.87	83.84	83.59	83.99	81.42	81.42	82.90	81.26	73.27	73.84	77.36	83.69	84.32	81.19 (+0.20)
DPO	81.73	82.73	83.75	83.83	84.26	81.51	81.02	82.91	82.02	74.74	75.28	77.98	83.80	85.05	81.47 (+0.48)
PPO	81.27	80.90	82.27	82.67	83.00	81.25	80.06	81.47	81.62	72.33	73.74	76.23	82.48	83.53	80.20 (-0.79)
GRPO	81.50	79.64	84.41	83.21	85.49	80.17	83.56	82.40	82.78	81.39	**84.48**	**85.50**	85.40	85.49	83.24 (+2.25)
SimLLM	83.28	84.43	85.97	85.33	85.98	83.32	82.44	84.37	83.53	79.65	82.51	84.26	85.86	**86.52**	84.10 (+3.11)
CultureSPA	81.28	82.90	83.16	81.91	82.44	81.31	82.17	81.77	81.88	73.61	77.10	80.27	82.27	82.55	81.04 (+0.05)
Anchor-only	82.23	78.39	81.82	81.52	80.57	79.65	79.27	80.56	81.92	79.15	72.82	74.14	81.71	82.12	79.71 (-1.28)
Causal-only	83.79	82.54	85.93	85.60	84.96	82.60	85.91	81.61	85.04	73.23	75.99	79.01	84.56	85.91	82.62 (+1.63)
ACE-Align	**84.95**	**85.91**	**87.34**	**87.37**	**87.05**	**84.73**	**86.69**	**84.89**	**87.30**	**85.03**	79.05	84.95	**87.35**	86.50	**85.65** (+4.66)
$G = 4$															
Base model	81.82	82.64	83.83	83.32	83.93	80.61	79.94	82.54	80.87	72.67	72.80	75.72	83.52	84.15	80.60
AdPrompt	82.57	82.72	83.38	83.26	83.49	81.13	80.65	82.70	80.89	73.23	73.13	76.57	83.18	83.72	80.76 (+0.16)
DPO	81.78	83.02	83.73	83.79	84.02	81.38	80.53	82.80	82.07	74.86	74.62	77.42	83.51	84.65	81.30 (+0.70)
PPO	81.39	81.23	82.18	82.72	82.77	81.21	79.84	81.30	81.58	72.37	73.64	76.28	82.28	83.34	80.15 (-0.45)
GRPO	81.04	80.01	83.94	82.65	85.01	79.69	82.72	82.23	82.16	81.49	**84.01**	**84.80**	84.78	84.77	82.81 (+2.21)
SimLLM	83.13	84.57	85.64	85.18	85.57	83.10	81.81	84.14	83.28	79.54	81.96	83.71	85.42	85.96	83.79 (+3.19)
CultureSPA	80.93	82.88	82.72	81.46	81.89	80.90	81.18	81.44	81.38	73.67	76.73	79.64	81.66	81.81	80.59 (-0.01)
Anchor-only	81.30	77.47	81.34	81.27	80.83	79.25	78.58	80.13	81.63	78.31	72.42	74.27	81.55	82.30	79.33 (-1.27)
Causal-only	83.76	82.43	85.75	85.41	84.70	82.45	85.11	81.48	84.86	73.35	75.60	78.53	84.14	85.40	82.36 (+1.76)
ACE-Align	**84.90**	**86.17**	**87.21**	**87.16**	**86.85**	**84.55**	**86.10**	**84.81**	**87.00**	**84.67**	78.53	84.77	**86.91**	**86.27**	**85.42** (+4.82)

Table 1: Cultural alignment scores under varying persona granularities $G \in \{1, 2, 3, 4\}$. Best results are shown in bold and second-best results are underlined.

option (AlKhamissi et al., 2024). **(3) SFT-based Methods:** We include *SimLLM*, *CultureSPA*, and the no-augmentation variant of *CultureLLM* as supervised fine-tuning baselines (Cao et al., 2025; Xu et al., 2025; Li et al., 2024). **(4) RL-based Methods:** We include DPO, PPO, and GRPO to test whether standard preference or policy optimization improves cultural alignment without explicit attribute-effect matching (Rafailov et al., 2023; Schulman et al., 2017; Shao et al., 2024). **(5) Ablation Variants:** Anchor-only is equivalent to the no-augmentation CultureLLM variant and optimizes only $\mathcal{L}_{\text{anchor}}$ (Eq. 12 with $\beta = 0$). Causal-only optimizes only $\mathcal{L}_{\text{CE}}$ (Eq. 12 with $\alpha = 0$). Together, they isolate the two objectives in ACE-Align.



(a) ISSP Base Scores by Country (b) ACE-Align Gains by Group (c) ISSP Group Scores and Gaps

Figure 3: **Within-ISSP geographic equity across persona granularities.** (a) Base-model country scores averaged across $G \in \{1, 2, 3, 4\}$, sorted by score and color-coded by geographic group. (b) Group-mean gains of ACE-Align over the base model. (c) Group-mean absolute scores for the base model and ACE-Align; shaded bands show the directional Global North–South gap, defined as mean(Global North) minus mean(Global South). ACE-Align reduces this gap at every granularity.

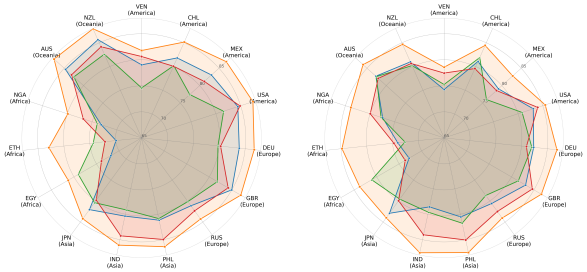
Evaluation Metric. Following Santurkar et al. (2023), we evaluate cultural alignment by comparing the model predicted opinion distribution to the reference human distribution on each multiple choice survey question. For each question $q \in Q$, let $D_1(q)$ and $D_2(q)$ denote two distributions over the answer choices (e.g., model vs. survey). To respect the ordinal structure of survey options, we use the 1-Wasserstein distance after embedding options into a 1D metric space (details below). We then define the alignment score S between D_1 and D_2 over a question set Q as

$$S = \frac{1}{|Q|} \sum_{q \in Q} \left(1 - \frac{\mathrm{WD}(D_1(q), D_2(q))}{N - 1} \right), \quad (13)$$

where N is the number of ordinal answer choices, and $N - 1$ is the maximum possible Wasserstein distance, ensuring $S \in [0, 1]$. For readability, all reported scores are scaled by 100.

4.2 Experimental Results

RQ1: Does ACE-Align consistently improve cultural alignment across varying persona granularities? Table 1 shows that ACE-Align achieves the highest average alignment scores across all persona granularities $G \in \{1, 2, 3, 4\}$. Compared to the base model, ACE-Align improves the average score by +4.62 at G=1, +4.51 at G=2, +4.66 at G=3, and +4.82 at G=4. Country-clustered paired tests in Appendix K further compare ACE-Align with all baselines in Table 1. Although performance generally reduces as G increases, higher granularity creates smaller persona subgroups and more compositional alignment constraints. ACE-Align nevertheless maintains robust gains across all granularities, indicating that its improvements persist from coarse to fine persona granularity. Appendix C further reports category-level alignment patterns across persona granularities.



(a) Holdout group 1 (b) Holdout group 2

Figure 4: —Base model — ACE-Align — Anchor Only — Causal Only Alignment on two held-out demographic combinations across cultural themes (Female, College Educated, Urban, Married) vs. (Male, Not College Educated, Rural, Not Married).

RQ2: How robust is ACE-Align under low-resource conditions? Low-resource conditions encompass (i) unseen persona compositions and (ii) low-coverage geographic regions.

Compositional Generalization. As demonstrated in Figure 4, ACE-Align maintains stable performance on persona profiles whose attribute combinations were withheld during training. This suggests the model captures reusable attribute-effect patterns rather than memorizing training instances.

Geographic Equity. Following the Brandt-line North–South division (Brandt et al., 1980), we group USA, DEU, GBR, JPN, AUS, and NZL as the Global North, and CHL, MEX, RUS, IND,

PHL, EGY, ETH, and NGA as the Global South. We classify Russia as a Global South emerging power (Jiang and Belli, 2025).

To avoid confounding geographic differences with the evaluation protocol, we compute the Global North–South gap separately within WVS and ISSP. For WVS, all 14 countries are evaluated using the same 80/20 held-out protocol. For ISSP, we compare the six Global North countries with the five covered Global South countries; EGY, ETH, and NGA are excluded because ISSP data are unavailable.

Figure 3 reports the ISSP-only comparison. ACE-Align improves the Global North group by 2.93, 2.69, 2.79, and 2.99 points and the Global South group by 4.69, 4.39, 4.44, and 4.57 points at $G \in \{1, 2, 3, 4\}$, respectively. The corresponding ISSP gap decreases from 2.70, 2.55, 2.49, and 2.37 points for the base model to 0.93, 0.84, 0.83, and 0.79 points for ACE-Align.

Appendix Figure 7 reports the separate WVS comparison: the gap decreases from 3.38, 3.45, 3.42, and 3.34 points to 1.23, 1.08, 1.02, and 1.09 points. Averaged across granularities, the gap shrinks from 3.40 to 1.11 points on WVS and from 2.53 to 0.85 points on ISSP. Appendix I shows that the conclusion remains unchanged when Russia is reassigned to the Global North.

RQ3: What mechanism drives the effectiveness of ACE-Align? To understand *why* ACE-Align works, we analyze the underlying attribute-effect dynamics of cultural attributes. We hypothesize that its effectiveness stems from two key capabilities: (1) correctly identifying that different cultural topics are governed by distinct demographic drivers, and (2) actively correcting specific types of attribute-effect misalignment.

Heterogeneity of Attribute Influence. Cultural values are not uniformly influenced by all demographics, and the dominant attribute varies substantially across topics and across countries. As shown in Figure 5, we visualize the survey-grounded attribute-effect magnitude of different attributes across WVS topics. For *Economic Values*, the dominant attribute is *Gender* in Australia and *Residence* in India. For *Perceptions of Migration*, it shifts from *Education* in Australia to *Residence* in India. Appendix D reports the same analysis for the remaining countries. This supports the design of ACE-Align, which models these variations through the attribute-effect matching loss ($\mathcal{L}_{\text{CE}}$).

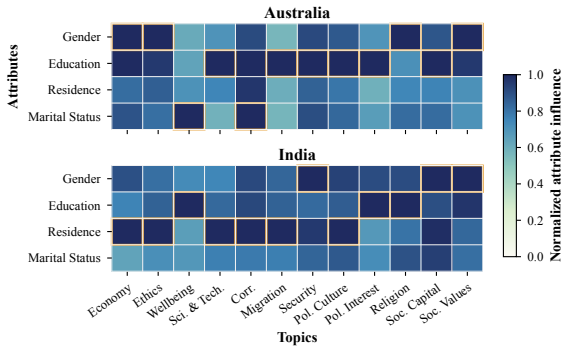


Figure 5: Heterogeneity of Attribute Influence for Australia and India. Different cultural topics and countries are dominated by different demographic attributes. Topic abbreviations are defined in Table 3.

Mitigating Stereotyping and Erasure. Finally, we analyze how the model aligns these attributes. We categorize the alignment status of each attribute-topic pair into four types based on the relationship between the model's estimated attribute effect (Δ_m) and the survey-grounded data effect (Δ_d). Table 2 summarizes the resulting taxonomy.

Type	Criterion	Interpretation
Flipped	$\text{sgn}(\Delta_m) \neq \text{sgn}(\Delta_d)$	Opposite trend
Stereotyping	$\lvert\Delta_m\rvert > \lvert\Delta_d\rvert + \epsilon$	Overestimated effect
Erasure	$\lvert\Delta_m\rvert < \lvert\Delta_d\rvert - \epsilon$	Underestimated effect
Aligned	$\lvert\Delta_m - \Delta_d\rvert \leq \epsilon$	Matched effect

Table 2: Taxonomy of attribute-effect alignment types, where ϵ denotes the tolerance.

Figure 6 shows that ACE-Align increases the share of *Aligned* effects across tolerance values ϵ. Because this category requires agreement in both direction and magnitude, it directly reflects whether controlled persona edits follow survey-grounded response shifts. Appendix E reports category-wise distributions for all countries. Together, these results show that ACE-Align improves both aggregate alignment and structural attribute-effect consistency.

5 Conclusion

We presented ACE-Align, a controlled persona-edit framework for cultural value alignment under varying persona granularities. ACE-Align matches model-side response shifts with survey-grounded attribute effects while anchoring absolute predictions to empirical response modes.

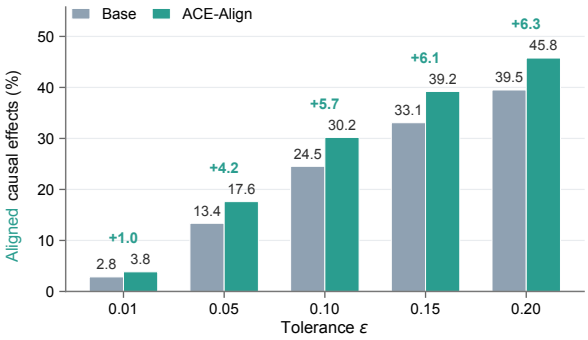


Figure 6: Aligned attribute-effect share across tolerance values ϵ (USA).

Across 14 countries and four persona granularities, ACE-Align improves cultural alignment. In within-survey comparisons, it reduces the average Global North–South alignment gap from 3.40 to 1.11 points on WVS and from 2.53 to 0.85 points on ISSP. These results show that modeling attribute-level response patterns can improve robustness to persona granularity and reduce geographic disparity in cultural alignment.

Limitations

There are still some practical limits to our study that we need to address. Because we rely on existing social surveys, we cannot represent every single culture or small community across the globe. The specific data available in these surveys also influenced our study, as we were limited to choosing only four primary traits: gender, education, residence, and marital status. Currently, we treat these traits as binary categories, such as urban versus rural or married versus not married. This binary approach is a simplified version of reality and does not capture the full complexity of how people actually identify themselves. Unobserved demographic factors and interactions among attributes may also affect survey responses, so controlled persona edits should be interpreted as observed-covariate contrasts rather than fully identified real-world causal effects. In the future, we want to look beyond these simple binary pairs and study more multivariate demographic factors. Moving toward these diverse and multi-layered factors will help our models reflect the true diversity of global cultures more fairly.

Ethics Statement

This work uses publicly available, de-identified survey datasets and does not collect new human-subject data. ACE-Align models population-level associations between demographic attributes and cultural-value responses. These associations should not be interpreted as deterministic descriptions of individuals or as fully identified real-world causal effects. Improper deployment may reinforce cultural stereotypes, oversimplify demographic groups, or disadvantage populations that are underrepresented in the surveys. Accordingly, the framework is intended for research and diagnostic purposes. Model outputs should be interpreted in context and with appropriate human oversight.

Acknowledgments

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) under Grant No. XDB0680302, the Director's Fund Project of the State Key Laboratory of AI Safety, and the Young Elite Scientists Sponsorship Program of the Beijing High Innovation Plan under Grant No. 20250924. We used GPT for polishing the text.

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

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*.

Willy Brandt and 1 others. 1980. *North South: a programme for survival; report of the independent commission on international development issues*. MIT, Cambridge, MA, US.

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.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello Piqueras, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. In *Proceedings of the first workshop on cross-cultural considerations in NLP (C3NLP)*, pages 53–67.

Isha Chaudhary, Qian Hu, Manoj Kumar, Morteza Ziyadi, Rahul Gupta, and Gagandeep Singh. 2025. Certifying counterfactual bias in llms. In *International Conference on Learning Representations*, volume 2025, pages 10494–10519.

Rongxin Chen, Yunfan Li, Yige Yuan, Bingbing Xu, and Huawei Shen. 2026. Multi-personality generation of llms at decoding-time. In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pages 58–68.

Rémy Decoupes, Roberto Interdonato, Mathieu Roche, Maguelonne Teisseire, and Sarah Valentin. 2025. Evaluation of geographical distortions in language models. *Machine Learning*, 114(12):1–32.

Bangde Du, Ziyi Ye, Zhijing Wu, Monika A Jankowska, Shuqi Zhu, Qingyao Ai, Yujia Zhou, and Yiqun Liu. 2025. Simvbg: Simulating individual values by backstory generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13104–13133.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bjorn Puranen, and 1 others. 2022. World values survey: Round seven–country-pooled datafile version 5.0.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Min Jiang and Luca Belli. 2025. Digital sovereignty in the brics countries.

Charles Kenny and Dev Patel. 2017. Gender laws, values, and outcomes: Evidence from the world values survey. *Center for Global Development Working Paper*, (452).

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

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

Sitiporn Sae Lim, Can Udomcharoenchaikit, Peerat Limkonchotiwat, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. Identifying and mitigating annotation bias in natural language understanding using causal mediation analysis. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11548–11563.

Vera Lomazzi and Daniel Seddig. 2020. Gender role attitudes in the international social survey programme: Cross-national comparability and relationships to cultural values. *Cross-Cultural Research*, 54(4):398–431.

Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C Treleaven, and Miguel Rodrigues Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Celia Serrano-Montilla, Luis M Lozano, Michael Bender, and Jose-Luis Padilla. 2020. Individual and societal risk factors of attitudes justifying intimate partner violence against women: a multilevel cross-sectional study. *BMJ open*, 10(12):e037993.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Nicholas Sukiennik, Chen Gao, Fengli Xu, and Yong Li. 2025. An evaluation of cultural value alignment in llm. *arXiv preprint arXiv:2504.08863*.

Zhouhao Sun, Li Du, Xiao Ding, Yixuan Ma, Yang Zhao, Kaitao Qiu, Ting Liu, and Bing Qin. 2024. Causal-guided active learning for debiasing large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14455–14469.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2022. On the machine learning of ethical judgments from

natural language. In *2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 769–779. 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.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Jia Wang, Ziyu Zhao, Tingjuntao Ni, and Zhongyu Wei. 2025. Sociobench: Modeling human behavior in sociological surveys with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26268–26300.

Junda Wu, Tong Yu, Xiang Chen, Haoliang Wang, Ryan Rossi, Sungchul Kim, Anup Rao, and Julian McAuley. 2024. Decot: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14073–14087.

Shaoyang Xu, Yongqi Leng, Linhao Yu, and Deyi Xiong. 2025. Self-pluralising culture alignment for large language models. 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 6859–6877.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Congzhi Zhang, Linhai Zhang, Jialong Wu, Yulan He, and Deyu Zhou. 2025. Causal prompting: Debiasing large language model prompting based on front-door adjustment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25842–25850.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17696–17706.

Xiaoyan Zhao, Juntao Jun, Yang Zhang, Wenjie Wang, Hong Cheng, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2026. Nextquill: Causal preference modeling for enhancing llm personalization. In *International Conference on Learning Representations*, volume 2026, pages 118745–118770.

A Details of Data Statistics

A.1 World Values Survey

Tables 3 and 4 summarize the WVS Wave 7 statistics used in this work, including culture topics, abbreviations, question counts, and country-level respondent counts after preprocessing.

Abbr.	Category	#Q
Soc. Values	Social Values, Norms, Stereotypes	44
Soc. Capital	Social Capital, Trust and Organizational Membership	36
Pol. Interest	Political Interest and Political Participation	36
Pol. Culture	Political Culture and Political Regimes	25
Ethics	Ethical Values	23
Security	Perceptions of Security	21
Religion	Religious Values	12
Wellbeing	Happiness and Wellbeing	11
Migration	Perceptions of Migration	10
Corr.	Perceptions of Corruption	9
Economy	Economic Values	6
Sci. & Tech.	Perceptions about Science and Technology	6
Demo.	Demographic and Socioeconomic Variables	1

Table 3: WVS topic categories, abbreviations, and question counts.

Continent	ID	Country	#Respondents
Americas	CHL	Chile	1000
	MEX	Mexico	1741
	USA	United States	2596
Europe	DEU	Germany	1528
	GBR	Great Britain	2609
	RUS	Russia	1810
Asia	IND	India	1692
	JPN	Japan	1353
	PHL	Philippines	1200
Africa	EGY	Egypt	1200
	ETH	Ethiopia	1230
	NGA	Nigeria	1237
Oceania	AUS	Australia	1813
	NZL	New Zealand	1057

Table 4: WVS respondents per country after preprocessing. Country identifiers follow ISO-3166-1 alpha-3. Continents are grouped into Americas, Europe, Asia, Africa, and Oceania.

A.2 International Social Survey Programme

Table 5 summarizes the ISSP dimensions used for evaluation, and Table 6 reports the corresponding country-level sample counts.

Dimension	#Questions
Work Orientations	66
Health	62
Citizenship	60
Role Of Government	59
National Identity	58
Social Networks	58
Religion	56
Family	50
Environment	49
Social Inequality	46

Table 5: ISSP dimensions and the number of survey questions per dimension.

B Persona Attribute Coding

Table 7 reports the survey-specific mappings used to harmonize the four demographic attributes and the binary coding used to define the direction of each persona-attribute effect. We treat $A = 1$ as the focal attribute value and $A = 0$ as the corresponding contrast value, so that estimated effect directions are comparable across countries, cultural topics, and persona granularities. Missing values, refusals, and "don't know" responses are treated as invalid and excluded before binarization.

C Category-level Results across Persona Granularities

We report category-level cultural alignment across countries under persona granularities $G \in \{1, 2, 3, 4\}$, as illustrated in Figure 10 to Figure 13. In each figure, each radar chart corresponds to one country, axes denote cultural categories, and curves compare the base model and ACE-Align.

D Heterogeneity of Attribute Influence

To complement the main analysis, we present attribute influence patterns for the remaining countries not shown in the main text, as illustrated in Figure 9. The figure visualizes the relative strength of demographic attributes across cultural topics, highlighting cross-country heterogeneity beyond the illustrative examples discussed earlier.

E Overall Comparison of Model Performance and Alignment

Figures 14–27 report category-wise attribute-effect alignment across countries. Each panel shows the share of attribute-effect pairs assigned to one alignment category across tolerance values ϵ. Colors indicate Base model, Aligned, Flipped, Stereotyping, and Erasure.

F Implementation Details

We perform parameter-efficient fine-tuning with Low-Rank Adaptation (LoRA) (Hu et al., 2022). We use rank $r = 8$, scaling factor $\alpha = 16$, and LoRA dropout of 0.05. Optimization is done with AdamW using a learning rate of 2×10^{-5}. We train on two NVIDIA A800 GPUs and use bfloat16 mixed precision.

We train for two epochs with a sequential objective schedule. In the first epoch, we optimize only the anchoring objective in Eq. 12 by setting $\alpha = 1.0$ and $\beta = 0.0$. In the second epoch, starting from the first-epoch checkpoint, we optimize only the effect-alignment objective by setting $\alpha = 0.0$ and $\beta = 1.0$. All other trained models follow the same optimization and hardware configuration unless stated otherwise.

G Additional Backbone Results

We additionally evaluate ACE-Align on Qwen3-8B, a similarly sized model from a different family, and Llama-3.1-70B-Instruct, a substantially larger model from the Llama family. Tables 8 and 9 report the complete country-level results. ACE-Align improves the macro-average over the corresponding base model at every persona granularity on both additional backbones.

Table 10 further compares ACE-Align with the evaluated baselines and ablations on Qwen3-8B. ACE-Align achieves the highest macro-average score at every granularity.

Finally, we quantify uncertainty in the improvement over each corresponding base model by resampling the 14 countries with replacement 10,000 times while retaining the paired ACE-Align/Base observations. As shown in Table 11, every paired 95% bootstrap confidence interval excludes zero.

G.1 RL Baseline Training Signals

For all RL-based baselines, we use the same backbone model, persona-conditioned prompt format,

Country	Citiz.	Envir.	Family	Health	Nat. Id.	Relig.	Govt.	Ineq.	Nets.	Work	**Total**
Australia	1432	1147	1612	1050	–	–	1267	1068	1317	1211	10104
Chile	1432	–	1564	–	–	1402	1416	1374	–	1433	8621
Germany	1718	1702	1766	1744	1717	1724	1689	1325	1701	1687	16773
Great Britain	1580	–	950	–	904	1552	1563	1724	1595	1793	11661
India	1209	1421	1660	1683	1530	–	1508	–	1510	1336	11857
Japan	1593	1491	1212	1453	1234	1466	1611	1473	1609	1573	14715
Mexico	–	–	1527	1001	1062	–	–	–	1002	1141	5733
New Zealand	–	993	–	1135	–	1334	1350	1210	1357	901	8280
Philippines	1200	1500	1200	1800	1200	1200	1200	4250	1200	1200	15950
Russia	1600	1583	1525	1597	1516	1583	1576	1597	1559	1596	15732
United States	1264	1847	1302	1146	1274	1175	1390	1852	1173	1477	13900

Table 6: ISSP sample counts by country and dimension. Dashes indicate that the corresponding module is not available for that country in our data.

Attribute	**WVS mapping**	**ISSP mapping**	$\mathbf{A = 1}$	$\mathbf{A = 0}$
Gender	Male vs. Female	SEX: 1 as Male and 2 as Female	Female	Male
Education	ISCED 0–3 as No College and ISCED 4–8 as College	DEGREE: 1–4 as No College and 5–6 as College. If only EDULEVEL is available, 0–4 and 5–8 are used	Not college educated	College educated
Residence	H_URBRURAL: Urban vs. Rural	URBRURAL: 1–3 as Urban and 4–5 as Rural	Rural	Urban
Marital status	Married or living together as married vs. divorced, separated, widowed, or single	MARITAL: 1–3 as Married and 4–6 as Not Married	Not married	Married

Table 7: Survey-specific demographic mappings and binary treatment coding used for effect-direction consistency.

training split, and evaluation pipeline as ACE-Align. The supervision signal is derived from the empirical survey response distribution. For each survey question x, let $\mathcal{Y}(x)$ denote the set of valid answer options, and let $p_{\text{data}}(y \mid x)$ be the empirical response distribution estimated from human survey responses. Model generations are parsed into a discrete answer option $y \in \mathcal{Y}(x)$ by extracting the first valid integer option from the response.

DPO preference pairs. For the DPO baseline, we construct preference pairs directly from the empirical survey distribution. For each survey question x, we enumerate pairs of valid answer options (y_i, y_j). If $p_{\text{data}}(y_i \mid x) > p_{\text{data}}(y_j \mid x)$, option y_i is treated as the preferred response and y_j as the rejected response. Pairs with equal empirical probabilities are omitted. Both preferred and rejected responses are formatted as the corresponding discrete option index, using the same answer-parsing convention as evaluation.

PPO reward. For the PPO baseline, we use a pointwise reward equal to the empirical probability of the selected answer:

$$r_{\text{PPO}}(x, y) = p_{\text{data}}(y \mid x). \tag{14}$$

If the model output cannot be parsed into a valid answer option, the reward is set to zero. This reward assigns larger scores to answers that are more common in the empirical survey distribution.

GRPO reward. For the GRPO baseline, rewards are computed using a group of sampled responses for the same question. Let $\mathcal{G}_x$ be the set of valid model responses sampled for question x in the current rollout group. We define the smoothed empirical model distribution as

$$q_{\text{batch}}(y \mid x) = \frac{c_x(y) + \lambda}{|\mathcal{G}_x| + \lambda|\mathcal{Y}(x)|}, \tag{15}$$

where $c_x(y)$ is the number of sampled responses equal to option y, and $\lambda = 0.5$ is the smoothing coefficient.

The reward for a valid response y is

$$r_{\text{GRPO}}(x, y) = \text{clip}_{[-2,2]}\Big[\log\left(p_{\text{data}}(y \mid x) + \epsilon\right) - \log\left(q_{\text{batch}}(y \mid x) + \epsilon\right)\Big], \tag{16}$$

where $\epsilon = 10^{-6}$. Invalid responses receive reward -2.

This reward is positive when an answer is undersampled by the model relative to the human distri-

G	Model	America			Europe			Asia			Africa			Oceania		Avg
		CHL	MEX	USA	DEU	GBR	RUS	IND	JPN	PHL	EGY	ETH	NGA	AUS	NZL	
$G=1$	Base model	80.78	80.24	81.13	81.17	81.24	79.01	78.03	78.76	79.30	67.63	71.88	73.78	81.42	82.19	78.33
	ACE-Align	**82.41**	**86.29**	**85.71**	**85.19**	**84.03**	**82.57**	**87.21**	**80.19**	**84.66**	**80.20**	**77.96**	**84.72**	**84.69**	**83.74**	**83.54** (+5.21)
$G=2$	Base model	80.75	79.98	81.19	81.43	81.69	79.31	78.34	79.58	79.72	69.24	72.32	74.28	81.57	82.40	78.70
	ACE-Align	**82.39**	**85.61**	**85.62**	**85.22**	**84.07**	**82.60**	**86.92**	**80.17**	**84.61**	**80.49**	**77.62**	**84.97**	**84.57**	**83.86**	**83.48** (+4.78)
$G=3$	Base model	80.81	80.02	81.17	81.67	81.92	79.47	78.36	80.03	79.87	70.55	72.47	74.33	81.64	82.39	78.91
	ACE-Align	**82.45**	**85.22**	**85.44**	**85.24**	**84.11**	**82.49**	**86.24**	**80.09**	**84.40**	**80.79**	**77.30**	**84.42**	**84.40**	**83.84**	**83.32** (+4.41)
$G=4$	Base model	80.85	80.08	81.04	81.68	81.99	79.51	78.34	80.27	79.88	72.09	72.07	74.29	81.56	82.32	79.00
	ACE-Align	**82.55**	**85.14**	**85.30**	**85.20**	**84.02**	**82.44**	**85.25**	79.95	**84.07**	**80.78**	**76.76**	**83.17**	**84.19**	**83.90**	**83.05** (+4.05)

Table 8: Country-level cultural alignment scores for Qwen3-8B. Parenthesized values report the macro-average improvement of ACE-Align over the corresponding base model.

G	Model	America			Europe			Asia			Africa			Oceania		Avg
		CHL	MEX	USA	DEU	GBR	RUS	IND	JPN	PHL	EGY	ETH	NGA	AUS	NZL	
$G=1$	Base model	82.00	80.76	84.24	84.05	84.07	82.00	80.75	83.32	81.55	76.53	75.43	76.82	83.14	84.18	81.35
	ACE-Align	**87.38**	**88.04**	**90.05**	**89.78**	**90.23**	**87.71**	**87.42**	**83.55**	**88.27**	**82.69**	**82.00**	**84.61**	**89.97**	**88.72**	**87.17** (+5.82)
$G=2$	Base model	81.99	80.30	83.46	83.92	83.73	81.74	80.37	83.23	81.56	76.97	74.89	76.85	82.93	83.81	81.12
	ACE-Align	**87.22**	**87.39**	**89.77**	**89.79**	**90.07**	**87.52**	**87.16**	**83.43**	**88.14**	**82.64**	**81.91**	**84.02**	**89.84**	**88.60**	**86.96** (+5.84)
$G=3$	Base model	81.84	80.07	82.94	83.71	83.49	81.41	80.05	83.00	81.59	77.16	74.36	76.62	82.66	83.34	80.87
	ACE-Align	**87.05**	**87.08**	**89.36**	**89.70**	**89.80**	**87.24**	**86.73**	**83.31**	**87.95**	**82.53**	**81.59**	**83.56**	**89.56**	**88.31**	**86.70** (+5.83)
$G=4$	Base model	81.74	80.18	82.26	83.32	83.16	81.03	79.55	82.69	81.42	77.49	73.74	76.39	82.24	82.80	80.57
	ACE-Align	**86.83**	**87.25**	**89.09**	**89.46**	**89.34**	**86.94**	**86.08**	**83.19**	**87.61**	**82.37**	**81.12**	**82.87**	**89.20**	**88.11**	**86.39** (+5.82)

Table 9: Country-level cultural alignment scores for Llama-3.1-70B-Instruct. Parenthesized values report the macro-average improvement of ACE-Align over the corresponding base model.

Method	$G=1$	$G=2$	$G=3$	$G=4$	Avg.
Base	78.33	78.70	78.91	79.00	78.74
AdPrompt	73.63	75.09	76.10	76.75	75.39
CultureSPA	76.56	76.77	76.77	76.81	76.73
DPO	77.84	78.25	78.50	78.57	78.29
PPO	76.16	76.48	76.74	76.75	76.53
GRPO	77.90	78.35	78.60	78.55	78.35
Anchor-only	79.51	79.59	79.37	79.02	79.37
Causal-only	82.11	81.99	81.70	81.29	81.77
ACE-Align	**83.54**	**83.48**	**83.32**	**83.05**	**83.35**

Table 10: Macro-average baseline comparison on Qwen3-8B.

Backbone	G	Improvement [95% CI]
Llama-3.1-8B	1	+4.62 [3.43, 6.11]
	2	+4.51 [3.23, 6.09]
	3	+4.66 [3.36, 6.23]
	4	+4.82 [3.56, 6.36]
Qwen3-8B	1	+5.21 [3.58, 7.09]
	2	+4.78 [3.22, 6.56]
	3	+4.41 [2.95, 6.06]
	4	+4.05 [2.77, 5.45]
Llama-3.1-70B	1	+5.83 [4.84, 6.58]
	2	+5.84 [4.85, 6.55]
	3	+5.82 [4.85, 6.52]
	4	+5.82 [4.86, 6.51]

Table 11: Macro-average improvements of ACE-Align over the corresponding base model, with paired country-bootstrap 95% confidence intervals.

bution and negative when it is over-sampled. Thus, unlike the PPO reward, it uses the current roll-out group to encourage closer agreement between sampled model responses and the empirical survey distribution.

H Supplementary WVS Held-out Evaluation

Table 12 reports WVS 80/20 held-out evaluation results for the non-African countries in the main study. This supplementary evaluation uses the same within-survey split protocol as the African countries in the main evaluation and serves as a robustness check for the ISSP/WVS protocol difference.

G	Model	America			Europe			Asia			Oceania		Avg
		CHL	MEX	USA	DEU	GBR	RUS	IND	JPN	PHL	AUS	NZL	
$G = 1$	Base model	77.85	77.83	80.43	78.68	81.45	78.32	76.80	77.53	76.78	80.97	80.66	78.85
	ACE-Align	**89.81**	**87.06**	**88.75**	**86.48**	**89.62**	**88.32**	**86.05**	**84.70**	**88.40**	**86.94**	**87.05**	**87.56** (+8.71)
$G = 2$	Base model	77.76	77.69	80.49	78.40	81.21	77.83	76.67	77.30	76.66	81.09	80.16	78.66
	ACE-Align	**89.62**	**87.74**	**88.59**	**86.22**	**90.05**	**88.02**	**86.29**	**84.39**	**88.74**	**87.45**	**86.94**	**87.64** (+8.98)
$G = 3$	Base model	77.30	77.58	80.08	77.95	80.98	77.36	76.43	76.91	76.50	80.83	79.72	78.33
	ACE-Align	**88.89**	**88.00**	**88.27**	**85.82**	**90.27**	**87.84**	**86.33**	**83.97**	**88.62**	**87.45**	**86.86**	**87.48** (+9.15)
$G = 4$	Base model	76.96	77.27	79.44	77.41	80.58	76.80	76.03	76.50	76.24	80.35	79.14	77.88
	ACE-Align	**88.73**	**87.89**	**87.94**	**85.04**	**90.25**	**87.46**	**86.11**	**83.98**	**88.08**	**87.20**	**86.82**	**87.23** (+9.35)

Table 12: Supplementary WVS 80/20 held-out evaluation for non-African countries. The average is computed over the 11 non-African countries in the main study.

I Additional Geographic Equity Results

Figure 7 reports the WVS-only Global North–South gaps. All 14 countries use the same WVS 80/20 held-out protocol. ACE-Align reduces the gap at every persona granularity under this within-survey comparison.

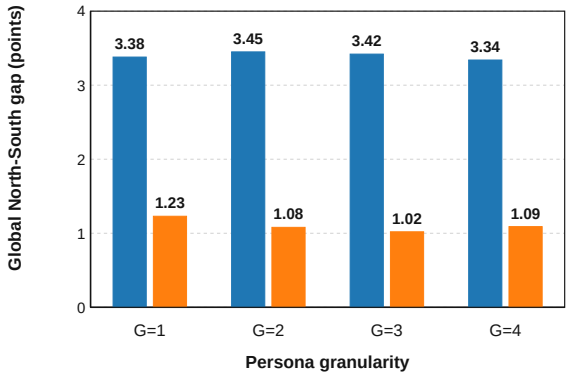


Figure 7: WVS-only Global North–South alignment gaps across persona granularities under the same 80/20 held-out protocol for all 14 countries. Each bar reports mean(Global North) minus mean(Global South). Lower values indicate a smaller geographic gap.

Our primary analysis follows the grouping in Figure 3, with Russia included in the Global South. Because Russia's contemporary classification remains debated, we conduct a sensitivity analysis that reassigns Russia to the Global North. Table 13 reports the Global North–South gaps averaged across $G \in \{1, 2, 3, 4\}$. ACE-Align reduces the gap under both groupings and within both evaluation protocols.

Grouping	WVS Base	WVS ACE	ISSP Base	ISSP ACE
RUS in Global South	3.40	**1.11**	2.53	**0.85**
RUS in Global North	3.33	**1.49**	1.89	**0.21**

Table 13: Sensitivity of the average Global North–South alignment gap to the geographic classification of Russia. Values are averaged across persona granularities; lower values indicate a smaller gap.

J Prompt Template

Figure 8 illustrates the persona prompting format used in all experiments.

You are a {GENDER} currently living in {RESIDENCE} in {COUNTRY}. You have {EDUCATION} , and you are {MARITAL_STATUS}.

Please answer the following survey question, and ensure that your choice reflects the persona described above. Do not add explanations or additional content. Respond with the number of the option you choose.

Question: {QUESTION_TEXT}

Options:
{KEY_1}: {OPTION_TEXT_1}
{KEY_2}: {OPTION_TEXT_2}
{KEY_3}: {OPTION_TEXT_3}
.....
{KEY_N}: {OPTION_TEXT_N}

Answer:

Figure 8: Persona-conditioned prompt template used for cultural alignment experiments.

K Statistical Significance

We assess statistical significance with country-clustered paired sign-flip tests. For each granularity and comparison, we first compute country-level paired score differences using the same aggregation protocol and 14-country set as Table 1. We compare ACE-Align with all nine baselines in Table 1: the base model, AdPrompt, DPO, PPO, GRPO, SimLLM, CultureSPA, Anchor-only, and Causal-only. We then run an exact two-sided sign-flip test over the 14 country-level differences and apply Holm correction over the 36 planned com-

parisons in Table 14. Confidence intervals are estimated by bootstrap resampling countries. All mean differences and bootstrap confidence intervals are positive. ACE-Align improves over all baselines except SimLLM and GRPO in all 14 countries at every granularity; for the stronger SimLLM and GRPO baselines, positive-country counts range from 11/14 to 13/14.

G	Comparison	Mean Δ	95% CI	p	Holm p
$G = 1$	vs. Base model	+4.62	[3.43, 6.11]	0.00012	0.00439
$G = 1$	vs. AdPrompt	+4.23	[3.13, 5.68]	0.00012	0.00439
$G = 1$	vs. DPO	+4.31	[3.34, 5.54]	0.00012	0.00439
$G = 1$	vs. PPO	+5.45	[4.31, 6.90]	0.00012	0.00439
$G = 1$	vs. GRPO	+2.02	[0.36, 3.36]	0.02808	0.08423
$G = 1$	vs. SimLLM	+1.38	[0.17, 2.49]	0.04407	0.08423
$G = 1$	vs. CultureSPA	+4.26	[3.31, 5.59]	0.00012	0.00439
$G = 1$	vs. Anchor-only	+5.57	[4.54, 6.55]	0.00012	0.00439
$G = 1$	vs. Causal-only	+3.19	[2.10, 4.79]	0.00012	0.00439
$G = 2$	vs. Base model	+4.51	[3.23, 6.09]	0.00012	0.00439
$G = 2$	vs. AdPrompt	+4.27	[3.11, 5.74]	0.00012	0.00439
$G = 2$	vs. DPO	+4.21	[3.19, 5.47]	0.00012	0.00439
$G = 2$	vs. PPO	+5.56	[4.41, 7.01]	0.00012	0.00439
$G = 2$	vs. GRPO	+2.19	[0.55, 3.54]	0.01941	0.08118
$G = 2$	vs. SimLLM	+1.42	[0.29, 2.50]	0.03162	0.08423
$G = 2$	vs. CultureSPA	+4.41	[3.47, 5.72]	0.00012	0.00439
$G = 2$	vs. Anchor-only	+5.79	[4.81, 6.83]	0.00012	0.00439
$G = 2$	vs. Causal-only	+3.05	[1.88, 4.70]	0.00012	0.00439
$G = 3$	vs. Base model	+4.66	[3.36, 6.23]	0.00012	0.00439
$G = 3$	vs. AdPrompt	+4.47	[3.29, 5.94]	0.00012	0.00439
$G = 3$	vs. DPO	+4.18	[3.16, 5.43]	0.00012	0.00439
$G = 3$	vs. PPO	+5.45	[4.33, 6.89]	0.00012	0.00439
$G = 3$	vs. GRPO	+2.41	[0.84, 3.68]	0.00952	0.06665
$G = 3$	vs. SimLLM	+1.55	[0.46, 2.58]	0.01624	0.08118
$G = 3$	vs. CultureSPA	+4.61	[3.69, 5.86]	0.00012	0.00439
$G = 3$	vs. Anchor-only	+5.95	[5.06, 6.95]	0.00012	0.00439
$G = 3$	vs. Causal-only	+3.03	[1.84, 4.67]	0.00012	0.00439
$G = 4$	vs. Base model	+4.82	[3.56, 6.36]	0.00012	0.00439
$G = 4$	vs. AdPrompt	+4.66	[3.51, 6.07]	0.00012	0.00439
$G = 4$	vs. DPO	+4.12	[3.13, 5.33]	0.00012	0.00439
$G = 4$	vs. PPO	+5.27	[4.20, 6.65]	0.00012	0.00439
$G = 4$	vs. GRPO	+2.61	[1.03, 3.87]	0.00720	0.05762
$G = 4$	vs. SimLLM	+1.64	[0.58, 2.62]	0.01038	0.06665
$G = 4$	vs. CultureSPA	+4.83	[3.92, 5.99]	0.00012	0.00439
$G = 4$	vs. Anchor-only	+6.09	[5.24, 7.06]	0.00012	0.00439
$G = 4$	vs. Causal-only	+3.07	[1.91, 4.63]	0.00012	0.00439

Table 14: Country-clustered paired significance tests for the planned ACE-Align comparisons. Mean Δ denotes the average country-level score difference in percentage points.

L Placebo Test for Causal Intervention Validity

We run a placebo intervention test to examine whether controlled prompt edits induce attribute-specific distributional shifts rather than generic sensitivity to superficial prompt changes. Using the base model in the United States setting, we compare three matched-pair interventions across all 10 survey categories and persona granularities $G \in \{1, 2, 3, 4\}$. The real intervention toggles *Education* (college educated vs. not college educated), while two placebo interventions toggle semantically irrelevant prompt fields: *respondent ID* (Respondent A vs. Respondent B) and *name initial* (names starting with A vs. B). For each intervention, the paired prompts differ only in the target field, and we measure the induced output-distribution shift with Earth Mover's Distance (EMD).

Across 14,826 matched pairs per intervention group, the real education intervention yields a mean ΔEMD of 9.44×10^{-2}, whereas the respondent-ID and name-initial placebos yield only 1.55×10^{-2} and 1.24×10^{-2}, respectively. As shown in Table 15, this 6–9$\times$ gap is consistent across all persona granularities. Welch's t-tests confirm that the gap between the real intervention and each placebo is significant across all comparisons ($t > 149$, $p < 0.001$), with large effect sizes (Cohen's $d > 1.7$). The small residual shifts under placebo interventions provide an estimate of baseline prompt-perturbation noise, while the substantially larger education shifts support the use of controlled persona edits as meaningful attribute interventions in our operational LLM setting.

G	Real ΔEMD	Placebo 1 ΔEMD	Placebo 2 ΔEMD	Real/P1	Real/P2	Cohen's d vs. P1/P2
1	11.86	1.70	1.33	7.0	8.9	1.86 / 1.94
2	10.53	1.74	1.37	6.1	7.7	1.79 / 1.87
3	9.37	1.58	1.26	5.9	7.5	1.75 / 1.83
4	8.42	1.34	1.10	6.3	7.6	1.72 / 1.79

Table 15: Placebo intervention test for prompt-based causal intervention validity. ΔEMD values are multiplied by 100 for readability. Placebo 1 toggles respondent ID, and Placebo 2 toggles name initial. All real-vs.-placebo comparisons have $p < 0.001$.

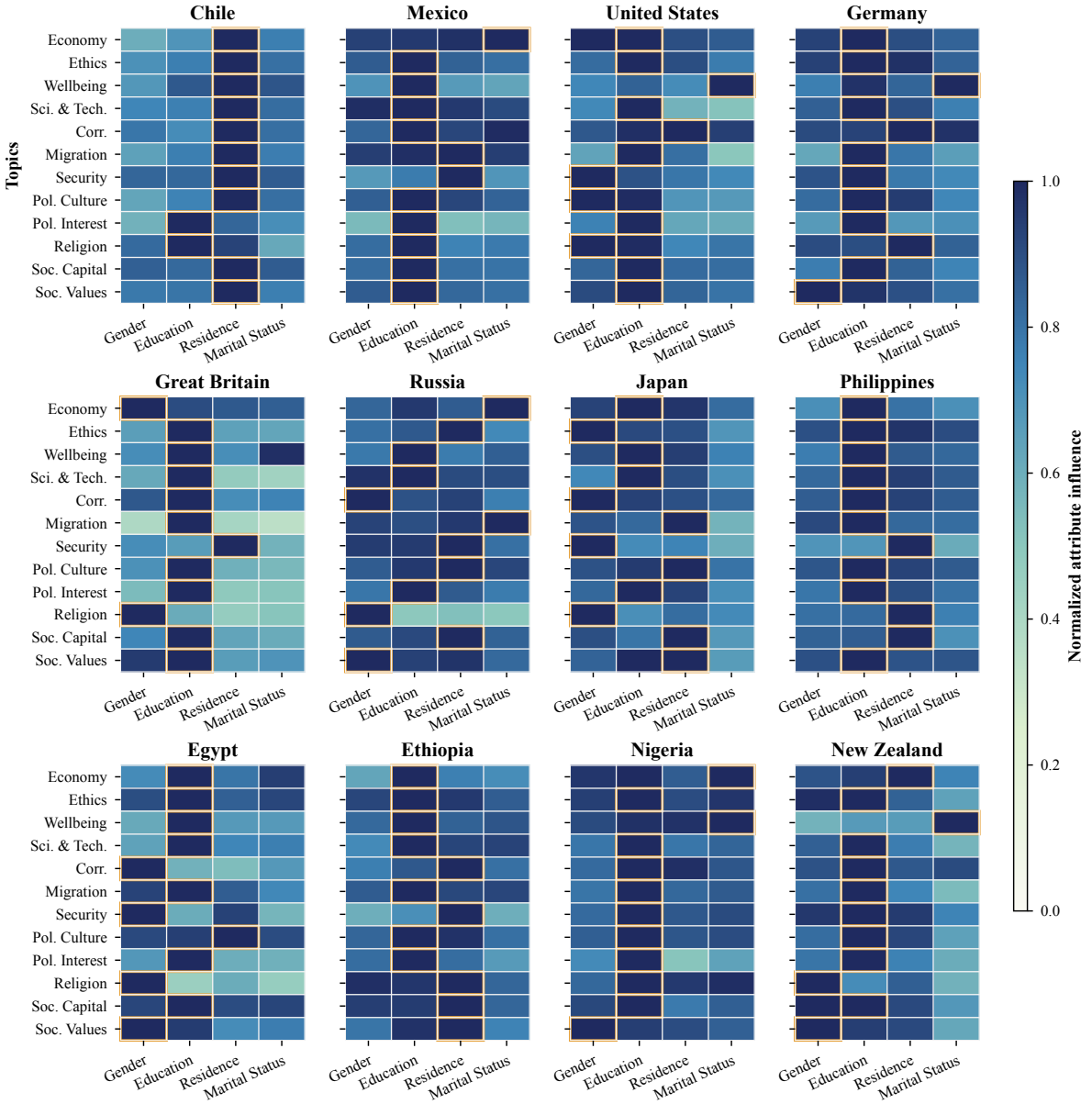


Figure 9: Heterogeneity of attribute influence across cultural topics for the remaining countries. Color intensity indicates the magnitude of the attribute effect. Topic abbreviations are defined in Table 3.

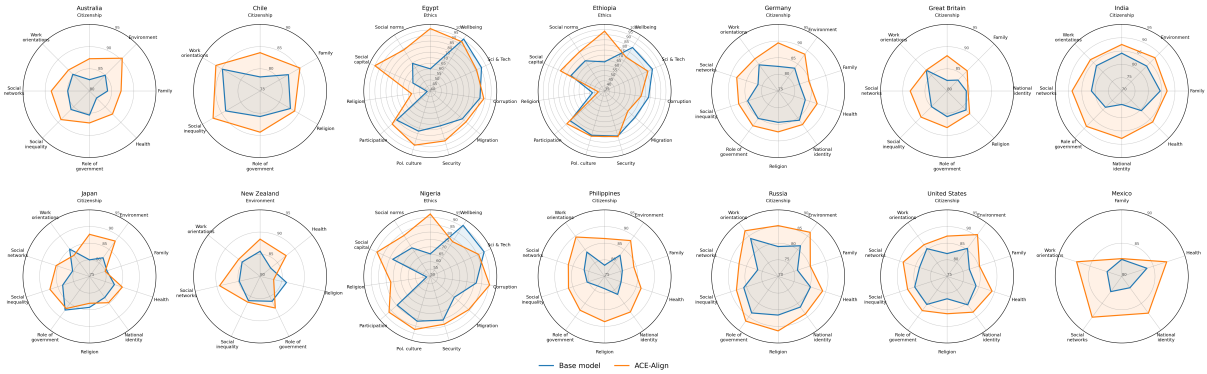


Figure 10: Category-level alignment under persona granularity $G = 1$.

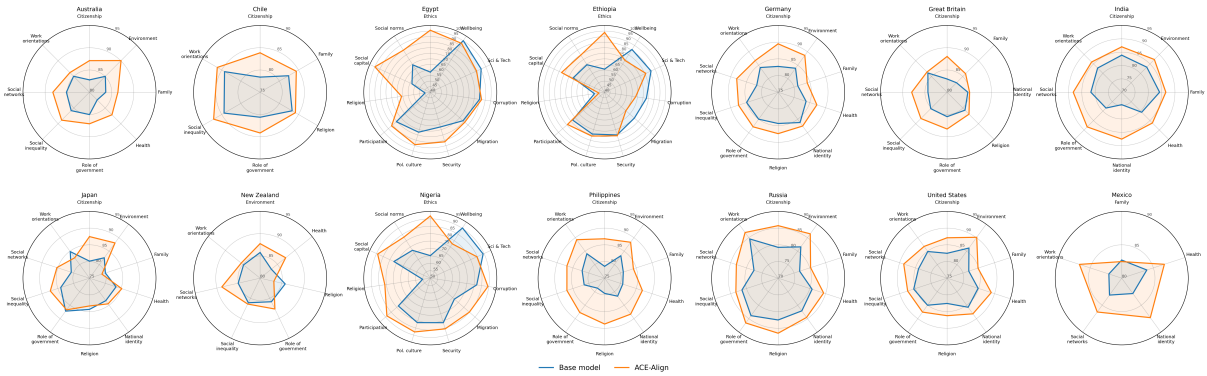


Figure 11: Category-level alignment under persona granularity $G = 2$.

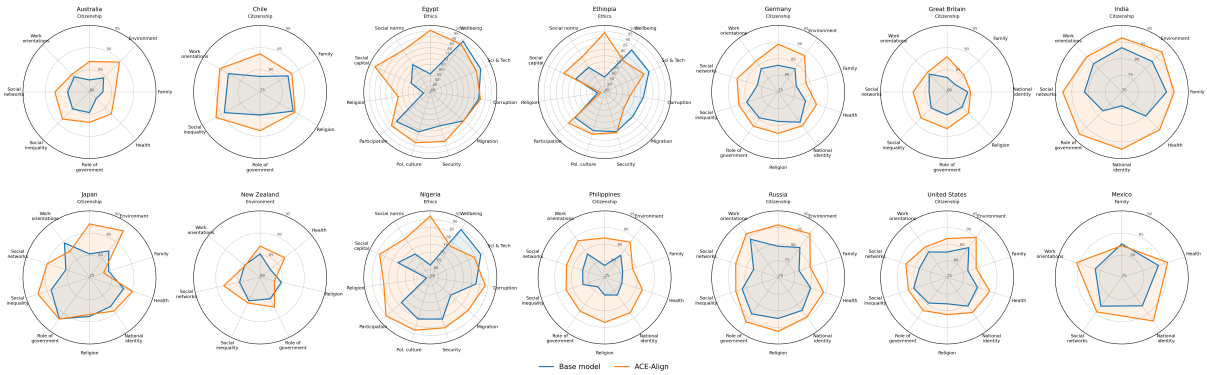


Figure 12: Category-level alignment under persona granularity $G = 3$.

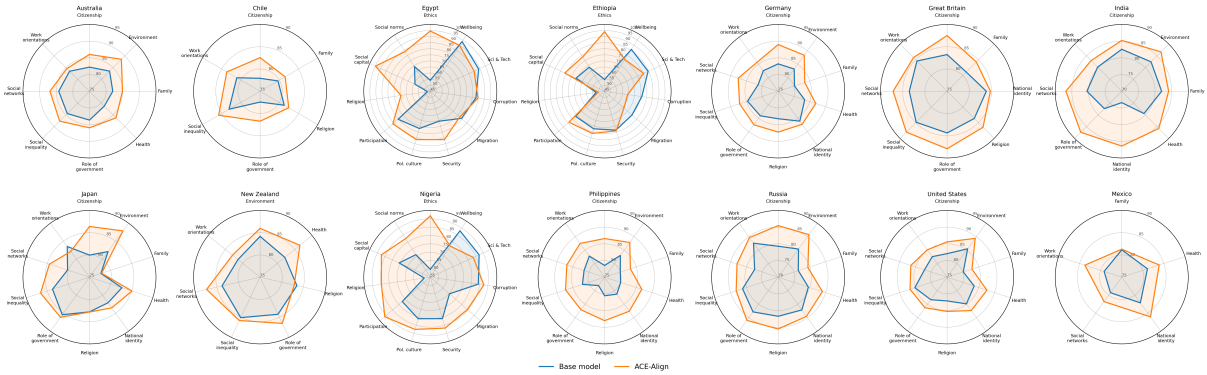


Figure 13: Category-level alignment under persona granularity $G = 4$.

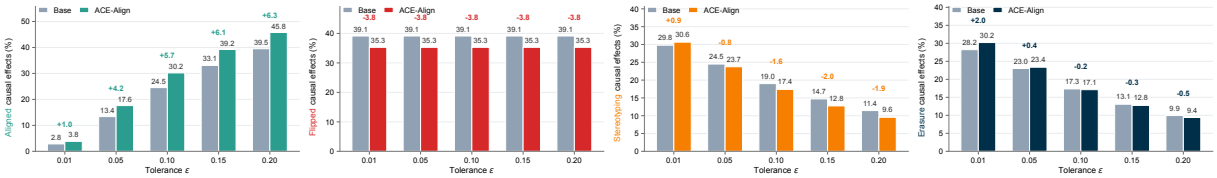


Figure 14: Category-wise attribute-effect shares across tolerance values ϵ (USA).

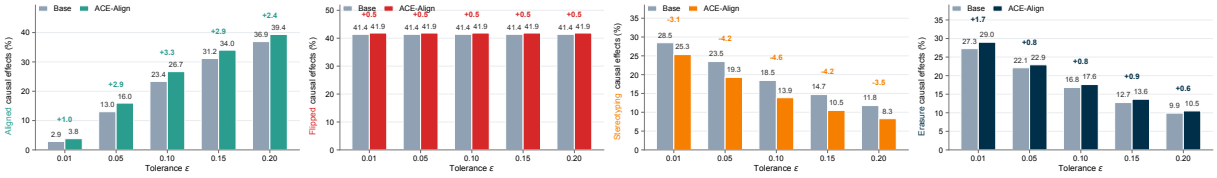


Figure 15: Category-wise attribute-effect shares across tolerance values ϵ (CHL).

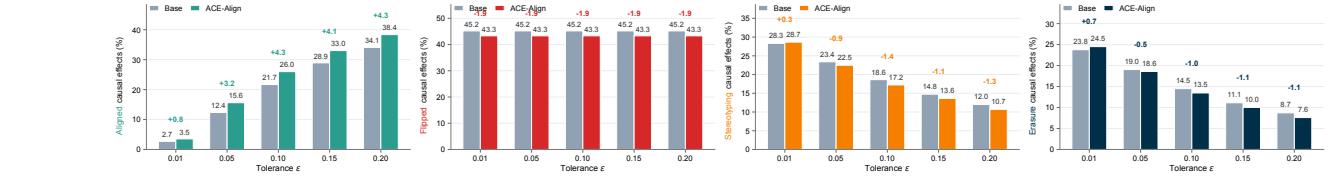


Figure 16: Category-wise attribute-effect shares across tolerance values ϵ (MEX).

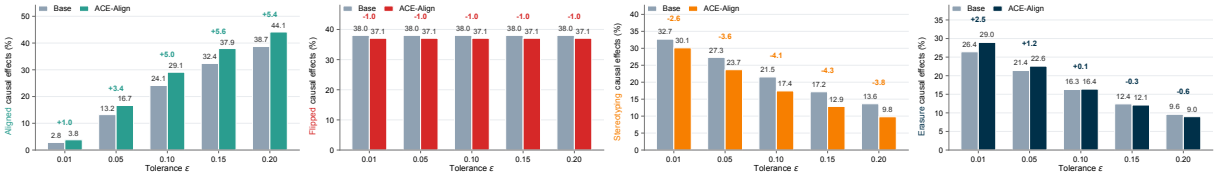


Figure 17: Category-wise attribute-effect shares across tolerance values ϵ (AUS).

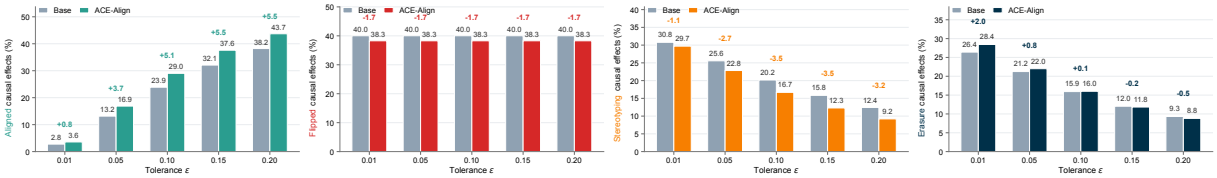


Figure 18: Category-wise attribute-effect shares across tolerance values ϵ (NZL).

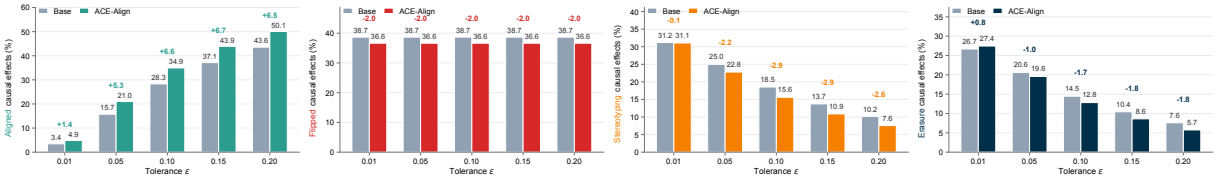


Figure 19: Category-wise attribute-effect shares across tolerance values ϵ (DEU).

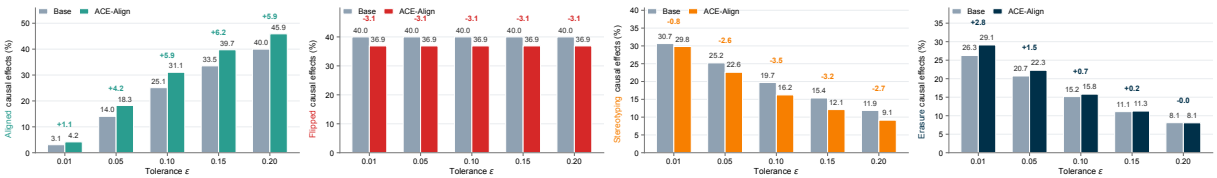


Figure 20: Category-wise attribute-effect shares across tolerance values ϵ (GBR).

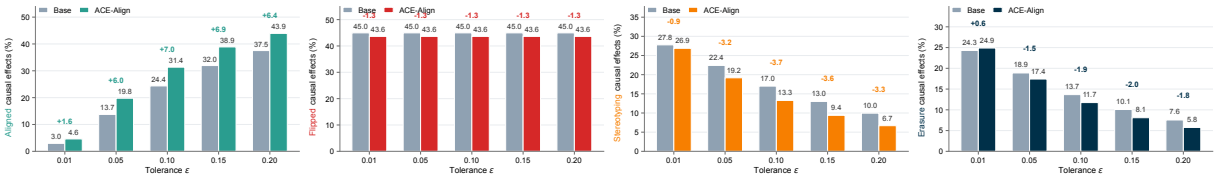


Figure 21: Category-wise attribute-effect shares across tolerance values ϵ (RUS).

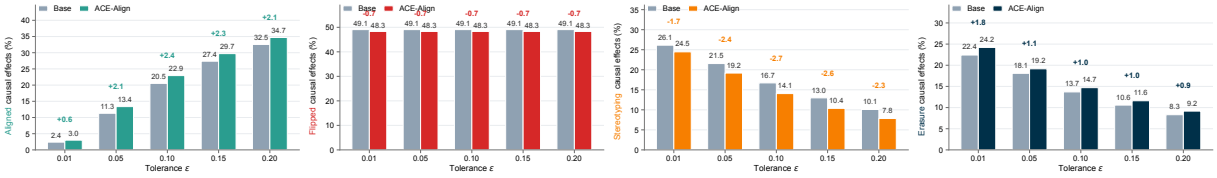


Figure 22: Category-wise attribute-effect shares across tolerance values ϵ (IND).

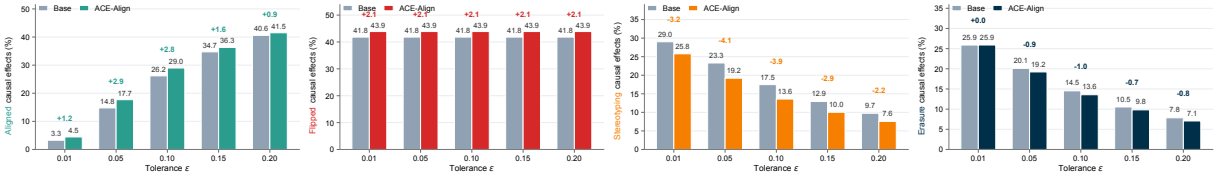


Figure 23: Category-wise attribute-effect shares across tolerance values ϵ (JPN).

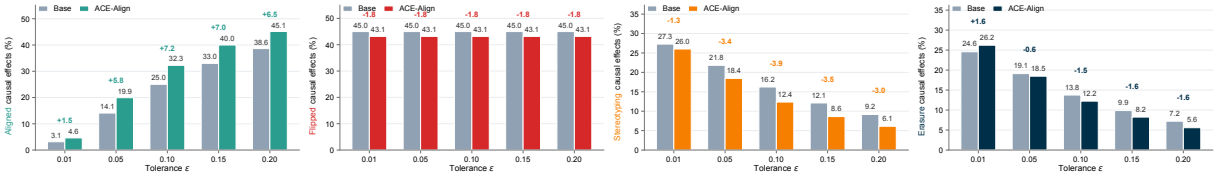


Figure 24: Category-wise attribute-effect shares across tolerance values ϵ (PHL).

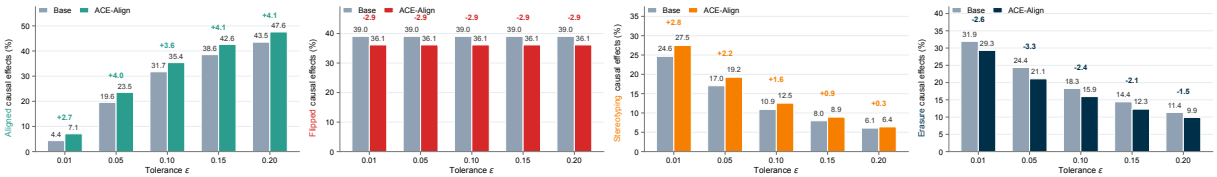


Figure 25: Category-wise attribute-effect shares across tolerance values ϵ (EGY).

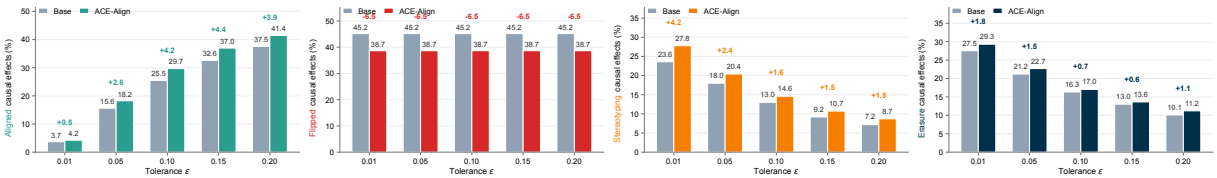


Figure 26: Category-wise attribute-effect shares across tolerance values ϵ (ETH).

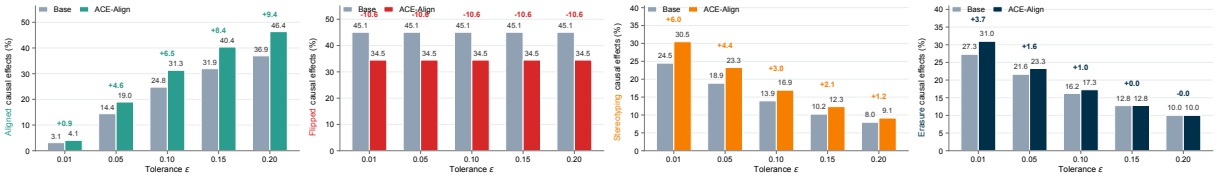


Figure 27: Category-wise attribute-effect shares across tolerance values ϵ (NGA).