Training-Free Cultural Alignment of Large Language Models via Persona Disagreement

Duy-Minh Dao-Sy[1,2,5,*] **Trung-Kiet Huynh**[1,2,5,*] **Tuan Nguyen**[3,*] **Chi-Nguyen Tran**[1,2,5]
Phu-Hoa Pham[1,2,5] **Phu Quy Nguyen Lam**[1,2,5] **The Anh Han**[4] **Long Tran-Thanh**[3,†]

[1]University of Science, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]University of Warwick, Coventry, United Kingdom
[4]Teesside University, Middlesbrough, United Kingdom
[5]CompactAI

{23122041,23122039,23122044,23122030,23122048}@student.hcmus.edu.vn
{tuan.nguyen.1,long.tran-thanh}@warwick.ac.uk, T.Han@tees.ac.uk

Abstract

Large language models (LLMs) increasingly shape decisions that turn on moral judgement, and moral preferences differ by country. The Moral Machine experiment collected 40 million judgments from 233 countries and territories and showed that these preferences follow culture, yet the preferences of LLMs sit closest to Western populations. A single deployment therefore serves every country with one society's intuitions. Existing fixes need per-country preference data, per-country reward models, or access to model internals. We work at inference time, with no weight updates, no labelled data and no internal access. We frame the task as country-conditioned value-preference calibration for tasks with a small, explicit decision space. Our key observation is that disagreement within a country is the steering signal. When culturally grounded personas agree, a correction toward their consensus can be trusted and refined. When they disagree, the refinement should shrink. DISCA represents each country as a panel of four persona prompts grounded in the World Values Survey. The panel's consensus sets a country-conditioned anchor, a loss-averse refinement beyond that anchor is kept only when two independent runs of it agree, and a running country prior pools the decisions made for one country. We evaluate on MultiTP across 20 countries and seven open-weight backbones from five model families. DISCA reduces cultural misalignment by 10 to 24% on the six backbones of 3.8B parameters or more, and with DISCA a 14B model is better aligned than a vanilla 70B model. On MultiTP, DISCA outperforms every training-free baseline we test and the standard aggregation rules applied to the same logits without the country prior. The same panel signal also improves social-norm classification on NormAd when the prompt gives only the country.

1 Introduction

Large language models now take part in decisions that turn on moral judgement: content moderation, clinical prioritisation, the behaviour of autonomous vehicles, and high-stakes recommendation. None of these decisions has a factual answer. Each one rests on a moral preference, and when a model is involved, the model's implicit preference becomes the system's preference.

[*]Equal contribution. [†]Lead senior author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

That preference cannot suit everyone, because moral preferences differ by country. The Moral Machine experiment collected 40 million judgments across 233 countries and territories on the same crash dilemmas and found that the answers are structured by culture [Awad et al., 2018]. LLM answers resemble those of Western, educated, industrialised, rich, and democratic (WEIRD) populations more than those of other countries [Henrich et al., 2010, Atari et al., 2023, Durmus et al., 2024]. They also stereotype the moral values of non-Western countries [Zewail et al., 2026], and post-training alignment adds further unintended shifts [Ryan et al., 2024]. The MultiTP benchmark [Jin et al., 2025] extends the Moral Machine protocol to 107 languages. It scores models against each country's Average Marginal Component Effects (AMCEs), which measure how much one attribute of a group changes the chance that the group is spared, and finds weak country-level alignment for most of the 19 LLMs it evaluates. A model that ignores this variation imposes one society's intuitions on every user it serves.

The existing fixes each ask for something the long tail of countries does not have. Fine-tuning and culture-aware adapters need culture-specific training data and a training run for every target population [Zhang et al., 2026, Yao et al., 2025]. Reward-guided decoding needs a trained reward model per country [Khanov et al., 2024, Mudgal et al., 2024]. Activation steering needs write access to model internals [Arditi et al., 2024, Turner et al., 2023], which black-box APIs do not expose. We therefore work in a *black-box, public-data-only regime*. The method runs at inference time, uses no per-country reward model and no labelled preference data, and reads only the log-probabilities of the decision tokens. Every open-weight backbone we evaluate exposes these scores, and so do the major commercial APIs that return `logprobs`. In this regime, cultural alignment can scale from a handful of curated countries to every country with public survey microdata, about 66 in the latest wave of the World Values Survey.

Our approach starts from a survey panel. A persona prompt does shift a model's moral decisions, and in fact shifts them more than the matching demographic differences shift human respondents [Kim et al., 2025]. A single persona therefore overshoots, and it reflects whichever person the prompt writer chose. A country's values are also not a single point. World Values Survey (WVS) microdata [Haerpfer et al., 2024] show that values inside one country vary with age cohort and social position [Inglehart and Welzel, 2005]. A pollster handles both problems with a panel: ask many respondents, report the average, and do not act on a difference smaller than the margin of error. Picture asking four local guides the way to the station. You walk in the direction their answers average out to. If the four agree closely, you also let them refine the last few steps. If they disagree, you stop at the average.

DISCA (Disagreement-Informed Steering for Cultural Alignment) does the same with a frozen model. For one country and one dilemma, it runs four persona prompts grounded in WVS microdata, three age cohorts and one country-wide aggregate, and records each persona's decision-token logit gap, the difference between the model's scores for the two answers. The panel then answers three questions:

Question	Answered by
Which way should the model move, and roughly how far?	The panel consensus, which forms a country-conditioned anchor.
Can that step be refined without hurting any persona?	A Prospect-Theory importance-sampling refinement [Kahneman and Tversky, 1979, Tversky and Kahneman, 1992] that penalises a change helping some personas while harming others.
Is the refinement reliable on this scenario?	A dual-pass gate: two independent runs must agree, and the refinement shrinks toward zero when they do not.

A fourth stage, a running country prior, pulls each decision toward the average of the decisions already made for the same country. A simple noise model makes the principle precise: the fraction of the consensus correction that minimises mean squared error falls as the panel variance rises (Proposition 1). DISCA follows that shape with a statistic it can compute at runtime.

We evaluate on 20 countries across four continents and on seven open-weight backbones from five model families, from 2B to 70B parameters. DISCA reduces cultural misalignment on MultiTP by 10 to 24% on the six backbones of 3.8B parameters or more, and it still improves the smallest 2B model. Phi-4 (14B) with DISCA is better aligned than a vanilla 4-bit Llama-3.3-70B, a model five times its size. On the Phi-4 MultiTP grid, DISCA outperforms every training-free baseline we test,

including activation steering and MC-Dropout. Given exactly the same MultiTP logits, it also beats probability averaging, logit consensus and majority vote applied without the country prior. The panel signal transfers beyond crash dilemmas and raises social-norm accuracy on NormAd [Rao et al., 2025] when the prompt gives only the country. The BLEnD factual cultural QA benchmark [Myung et al., 2024] marks the scope boundary: DISCA steers values, where the decision reduces to a scalar logit gap, but not factual recall, where the decision is one token in a large vocabulary. Our contributions are:

- **Method.** An inference-time cultural alignment method that needs no weight updates, no per-country reward models and no white-box access.
- **Principle.** A formal account of disagreement-driven shrinkage, with the disagreement between two independent runs used as a reliability signal.
- **Evidence.** An evaluation on 20 countries, four continents and seven open-weight backbones on binary moral dilemmas, with same-logit aggregation controls and transfer tests on social norms, survey opinions and factual QA.

2 Related Work

Benchmarks and cultural evaluation. LLM alignment to the country-specific AMCEs of MultiTP [Jin et al., 2025, Awad et al., 2018] varies widely across model families [Takemoto, 2024, bin Ahmad and Takemoto, 2025]. Model answers resemble WEIRD respondents most [Atari et al., 2023, Durmus et al., 2024], and RLHF adds unintended cultural shifts [Ryan et al., 2024]. Kim et al. [2025] show that sociodemographic personas shift LLM moral decisions more than the matching demographic differences shift humans, most strongly for political personas. This finding motivates our design. DISCA does not use personas to steer toward one viewpoint. It uses them to *measure* within-country disagreement and turns that measurement into a bounded correction.

Inference-time alignment. Pluralistic alignment can be steered at inference time by conditioning on a target group [Sorensen et al., 2024], but each existing approach gives up one of the constraints we target. Activation steering [Arditi et al., 2024, Turner et al., 2023] needs white-box access, reward-guided decoding [Khanov et al., 2024, Mudgal et al., 2024] needs per-country reward models, and culture-aware adapters [Zhang et al., 2026, Yao et al., 2025] need per-culture training. Prospect-theoretic alignment [Ethayarajh et al., 2024] uses the same value shape we adopt to train a policy. DISCA applies it at test time, per scenario, to aggregate heterogeneous personas.

Panels, consensus and ensembles. Self-consistency [Wang et al., 2023], LLM debate [Du et al., 2024, Liang et al., 2024] and standard ensembles usually average, vote or adjudicate over diverse outputs. DISCA uses the *spread* of persona logit gaps to decide how much refinement beyond the consensus to keep, and Proposition 1 shows that the error-minimising shrinkage of the consensus correction is a closed-form function of within-panel variance. Standard calibration, such as temperature scaling and MC-Dropout [Gal and Ghahramani, 2016, Kwon et al., 2026], changes confidence in a general way, without reference to any population.

3 Disagreement-Informed Steering for Cultural Alignment

We build DISCA in three steps. §3.1 fixes the setting. §3.2 states the principle the controller follows: under a simple model of the panel, the fraction of the consensus correction that minimises error falls as the panel disagrees. That fraction cannot be computed from what the panel supplies. §3.3 therefore builds a controller of the same shape from quantities available at runtime: a consensus anchor, a loss-averse refinement around it, a gate that keeps the refinement only when two independent runs of it agree, and a running country prior (Figure 1).

3.1 Setup

We consider a frozen language model f_θ on a forced-choice task. Each input x asks the model to choose between two options through decision tokens A and B. The model's logits at the decision position are $z(x) = [z_a, z_b]$, and the decision gap is $\delta(x) = z_b - z_a$. For a target country c we assume

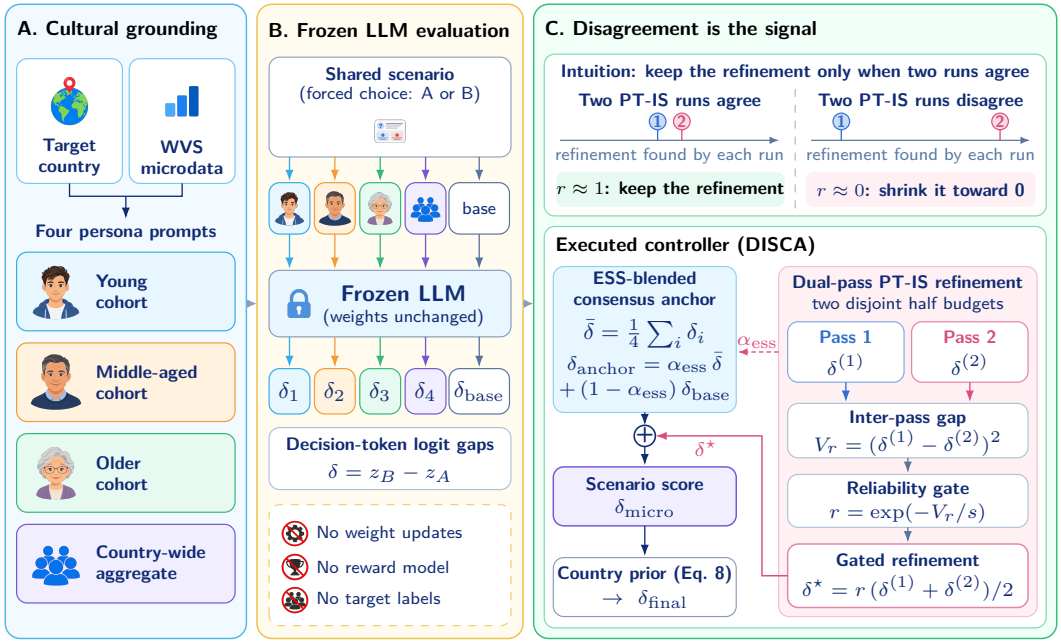


Figure 1: **DISCA overview.** **(A)** WVS microdata instantiate four persona prompts for country c: three age cohorts and a country-wide aggregate. **(B)** The personas and an unconditioned base prompt run through the frozen model on one scenario, giving decision-token logit gaps. **(C)** An anchor blends the persona consensus with the base gap, and two Prospect-Theory importance sampling (PT-IS) passes give a refinement $\delta^\star$ that the reliability gate shrinks when the passes disagree. The anchor plus the gated refinement gives the scenario score δ_{micro} (Eq. 7), and the running country prior (Eq. 8) blends it with the country's earlier scores into δ_{final}. Pseudocode in App. A1.

score-level access to f_θ and a public description of the population of c, here World Values Survey microdata [Haerpfer et al., 2024]. We assume no per-country preference labels, no reward model and no write access to θ. The goal is a per-scenario adjusted gap $\delta_{\text{final}}(x, c)$ whose decisions are closer to what the population of country c would choose, measured against a country-level reference vector (§4).

3.2 The disagreement principle

Evaluate the same scenario under N culturally grounded prompts for the same target country. Each prompt, or agent, produces a logit gap δ_i, a noisy estimate of what the population-level preference of country c implies for this scenario. If the N gaps cluster tightly, their average is a reliable target to move toward. If they are spread out, the average is a compromise between views the panel itself does not share, and moving all the way to it is a gamble. The principle is that the more the panel disagrees, the less of the correction should be applied. We make it precise with the simplest model that captures it. Let δ_h denote the unobserved gap consistent with the population-level preference of country c on that scenario, and model the agents as

$$\delta_i = \delta_h + \eta_i, \qquad \mathbb{E}[\eta_i] = 0, \qquad \mathrm{Var}(\eta_i) = \tau^2,$$

with the η_i independent and identically distributed. This is a simplification, since all agents share f_θ and the scenario. Let δ_{base} be the gap that the unconditioned base prompt, without persona or country information, gives on the same scenario. We define the consensus, the within-panel variance, the consensus adjustment and the oracle adjustment as

$$\bar{\delta} = \frac{1}{N}\sum_{i=1}^{N}\delta_i, \qquad D^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\delta_i - \bar{\delta})^2, \qquad \Delta = \bar{\delta} - \delta_{\text{base}}, \qquad \Delta_h = \delta_h - \delta_{\text{base}}.$$

A scaled correction $\hat{\Delta}(\gamma) = \gamma\Delta$ with $\gamma \in [0, 1]$ applies a fraction γ of the consensus adjustment, and we ask which fraction minimises

$$\mathrm{MSE}(\gamma) = \mathbb{E}\big[(\hat{\Delta}(\gamma) - \Delta_h)^2\big].$$

Proposition 1 (Variance-aware shrinkage; proof in App. A2)**.** *Under the agent model above:*

(i) D^2 *is unbiased for* τ^2*, and* $\mathrm{Var}(\Delta) = \tau^2/N$.

(ii) *Among estimators* $\hat{\Delta}(\gamma) = \gamma\Delta$ *with* $\gamma \in [0,1]$*, the MSE-minimising one has*

$$\gamma^\star = \frac{\Delta_h^2}{\Delta_h^2 + \tau^2/N},$$

which lies in $[0,1]$ *and decreases monotonically in* τ^2.

Part (ii) is the principle in closed form: the optimal fraction of the consensus step falls as the panel variance rises and never exceeds one. Part (i) says that the panel itself estimates τ^2 without bias through D^2. Two facts stop us from using $\gamma^\star$ directly. First, it depends on Δ_h, which nothing in the panel estimates. Second, with $N{=}4$ personas D^2 is unbiased but noisy on any single scenario: Theorem 3 (App. A3.2) bounds its deviation from τ^2, and with so few personas the bound is loose. The proposition therefore fixes the shape a controller should have: bounded, decreasing in disagreement, and smooth.

3.3 The DISCA estimator

We use $N{=}4$ personas per country: young, middle-aged and older cohorts, plus a country-wide aggregate. Each is a native-language system prompt built from cohort- and country-level aggregates of the WVS cultural profile of Greco et al. [2026] (App. A14). The three cohorts cover the age axis along which much of the within-country variation runs [Inglehart and Welzel, 2005], and the aggregate steadies the panel against cohort-specific noise. Under the noise model, the precision gained from one more persona decays as $\Theta(N^{-2})$ (Corollary 1). A sweep over $N \in \{2,3,4,5,6\}$ on a three-country panel finds that $N{=}4$ gives the lowest macro MIS (Table 24): $N{=}2$ leaves the consensus under-determined, and $N{\geq}5$ repeats demographic coverage without adding new variance.

Every gap is formed on a fixed scale. The two raw decision-token logits are first divided by a per-attribute logit temperature T_c (Table 5), so $\delta_i = (z_{b,i} - z_{a,i})/T_c$. To remove the order bias of forced-choice decoding, every scenario is evaluated under both the (A,B) and the (B,A) ordering, and the persona's gap is the order-symmetrised $\delta_i = \frac{1}{2}(\delta_i^{(AB)} - \delta_i^{(BA)})$. The base gap δ_{base} is formed the same way. The decision temperature T_{dec} enters only once, at the end.

Loss-averse importance sampling. The plain consensus adjustment can hide a cohort that strongly disagrees. Three cohorts mildly in favour and one strongly against average out to "mildly in favour", and the model then moves in a direction one cohort rejects. We frame this as a bargaining-inspired aggregation among the cohorts [Nash, 1950, Kalai and Smorodinsky, 1975], each a loss-averse player. We draw K candidate perturbations $\epsilon_k \sim \mathcal{N}(0,\sigma^2)$ around the consensus, giving candidate states $\tilde{\delta}_k = \bar{\delta} + \epsilon_k$. The proposal scale follows the panel,

$$\sigma = \max\{\mathrm{sd}(\delta_1,\ldots,\delta_N),\, 0.3\},$$

so the search is wider when the personas are further apart and never collapses when they agree. Each candidate is scored by how much closer it lands to each cohort gap and to the consensus than the base prompt did:

$$g_{i,k} = |\delta_{\text{base}} - \delta_i| - |\tilde{\delta}_k - \delta_i|, \qquad g_{\text{cons},k} = |\delta_{\text{base}} - \bar{\delta}| - |\tilde{\delta}_k - \bar{\delta}|. \tag{1}$$

A loss to any cohort should count more than a same-sized gain elsewhere, and gains beyond what is already aligned should count less. The canonical rule with both properties is the Kahneman-Tversky value function [Kahneman and Tversky, 1979, Tversky and Kahneman, 1992],

$$v(z) = \begin{cases} z^\alpha & z \geq 0, \\ -\kappa\,(-z)^\alpha & z < 0, \end{cases} \tag{2}$$

with curvature $\alpha \in (0,1]$ and loss aversion $\kappa \geq 1$. We blend a utilitarian verdict, the average utility of the loss-averse players, with a collective verdict against the consensus direction:

$$U_{\text{total}}(\epsilon_k) = (1-\lambda_{\text{coop}})\,\frac{1}{N}\sum_{i=1}^{N} v\Big(\frac{g_{i,k}}{\sigma}\Big) + \lambda_{\text{coop}}\, v\Big(\frac{g_{\text{cons},k}}{\sigma}\Big). \tag{3}$$

The weight $\lambda_{\text{coop}} \in [0, 1]$ sets the balance, from purely utilitarian at 0 to purely collective at 1. We use $\lambda_{\text{coop}} = 0.7$ (sweep in App. A13). Dividing by σ scores each gain in units of the perturbation scale, so σ and the value-function shape (α, κ) can be set independently.

Returning the highest-utility candidate is unsafe. The utility of each candidate is noisy, and an argmax over K noisy scores picks the candidate that drew favourable noise. We therefore aggregate all candidates with weights that grow with their utility,

$$\delta_{\text{PT-IS}} = \frac{\sum_{k=1}^{K} w_k \, \epsilon_k}{\sum_{k=1}^{K} w_k}, \qquad w_k \propto \exp\big(U_{\text{total}}(\epsilon_k)/\eta\big), \tag{4}$$

where $\eta = 0.5$ is a softmax temperature. This is a path-integral-style exponentially weighted update for choosing a correction under a noisy utility [Williams et al., 2018, Levine, 2018]: the best candidate dominates, and near-winners contribute too. We call the procedure *loss-averse importance sampling* (PT-IS). Its output is a convex combination of the draws, so it is bounded by the largest draw whatever the backbone does (Theorem 2). If the importance weights collapse, so that their normalised effective sample size falls below the threshold ρ_{eff} (Table 21), the pass returns zero.

Dual-pass reliability gate. A single PT-IS run cannot tell a backed direction from a contested one. Suppose two cohorts prefer a gap of $+0.5$ and two prefer -0.5. Loss aversion makes PT-IS return a small compromise, say $+0.05$, because pushing either way would hurt half the cohorts, and applying it nudges the model in a direction no cohort asked for. With a finite budget, each run also drifts toward whichever camp its samples happen to favour, so the refinement depends on the draw more than on the panel. More samples do not fix this. They make the compromise more precise without saying whether it deserves full strength. We could gate on D^2, but a four-sample variance is too noisy to gate on scenario by scenario. Instead we measure directly whether the refinement is repeatable. We run PT-IS twice on the same scenario, on disjoint budgets of $K_{\text{half}} = K/2$ draws, and compute the squared inter-run gap and the reliability weight

$$V_r = \big(\delta^{(1)}_{\text{PT-IS}} - \delta^{(2)}_{\text{PT-IS}}\big)^2, \qquad r = \exp(-V_r/s) \in (0, 1], \tag{5}$$

with bandwidth $s > 0$. V_r is zero when the runs agree and grows as they diverge. It has the form of the half-sample variance estimator of survey statistics [Mahalanobis, 1946, McCarthy, 1969, Wolter, 2007], which splits one sample in two and reads the variance from the squared gap between the halves. The exponential is the simplest function with the shape the proposition asks for: bounded in $(0, 1]$, decreasing in disagreement, and smooth. The retained refinement is the gated average of the two runs,

$$\delta^\star = r \cdot \frac{\delta^{(1)}_{\text{PT-IS}} + \delta^{(2)}_{\text{PT-IS}}}{2}. \tag{6}$$

The two half-runs keep total compute equal to one full PT-IS run. App. A9.2 relates panel disagreement to the size of the applied correction.

ESS anchor. The anchor blends the consensus with the base gap. Let $\overline{\text{ESS}}_m \in (0, 1]$ be the normalised effective sample size of the importance weights in pass m, a standard check of whether a weighted sample has degenerated, and let $\alpha_{\text{ess}} = \min\{1, \max\{\rho_{\text{eff}}, \min_m \overline{\text{ESS}}_m\}\}$. The scenario score is

$$\delta_{\text{micro}} = \alpha_{\text{ess}} \, \bar{\delta} + (1 - \alpha_{\text{ess}}) \, \delta_{\text{base}} + \delta^\star. \tag{7}$$

When the importance weights collapse onto a few candidates, α_{ess} falls and the anchor leans toward the base gap. When the two passes disagree and r approaches zero, the refinement vanishes and the score falls back to its anchor, not to δ_{base}.

Country-level running prior. The last stage pools information across the scenarios of one country. The controller keeps a running country score d_c and emits

$$\delta_{\text{final}} = a_t \, d_c + (1 - a_t) \, \delta_{\text{micro}}, \qquad a_t = \begin{cases} 0 & t < t_0, \\ 1 - e^{-(t - t_0)/\tau} & t \geq t_0, \end{cases} \tag{8}$$

after which it updates the exponential moving average $d_c \leftarrow (1-\omega) \, d_c + \omega \, \delta_{\text{micro}}$ with ω=0.1. Here t counts the scenarios already processed for c, t_0=50 is a warm-up, and τ=100 sets how fast the prior takes over. The decision is $p = \sigma(\delta_{\text{final}}/T_{\text{dec}})$. The prior uses only the controller's own earlier

outputs, and no human target or label enters it. It is, however, transductive: each decision depends on the earlier test scenarios of the same country and on their order, and the running score pools all six moral attributes. On the 310-scenario panel a_t reaches 0.93 by the last scenario. The anchor and the prior were part of every MultiTP run we report.

Beyond two decision tokens. The controller needs one score per answer option under each view, not specifically two tokens. With more options, the option scores are centred to sum to zero, the persona mean is the anchor, and the two PT-IS passes and the gate act in that higher-dimensional space. This multi-option form, used for NormAd, has no country prior (App. A11).

4 Experimental Results

We test DISCA on MultiTP [Jin et al., 2025], the multilingual extension of the Moral Machine trolley problem [Awad et al., 2018]. Each scenario shows two groups of people that an autonomous vehicle could spare, and the model chooses one by emitting a single decision token, A or B. Every scenario isolates one of six moral attributes (species, gender, age, fitness, social value, and the number of lives saved) and matches the two groups on everything else. An age scenario, for example, pits a young group against an older one, so the rate at which the model spares the younger group, averaged over many such scenarios, is its preference on age. Following the Moral Machine convention, decision token B is always the side the AMCE measures toward. This conjoint-analysis convention fixes the sign of the preference and makes no claim about which side ought to be saved. For vanilla decoding the sparing probability under decoding temperature T_{dec} is $p_{\text{spare}}(x) = \sigma(\delta(x)/T_{\text{dec}})$, and DISCA replaces $\delta(x)$ with δ_{final}. Averaging the sparing probability over the scenarios for one attribute and one country gives that country's AMCE:

$$\hat{m}_d^{(c)} = \frac{1}{|\mathcal{S}_d^{(c)}|} \sum_{x \in \mathcal{S}_d^{(c)}} p_{\text{spare}}(x). \tag{9}$$

Stacking the six attributes gives a vector $\hat{m}^{(c)}$, which we compare to the human AMCE vector $h^{(c)}$ aggregated from real Moral Machine judgments [Awad et al., 2018, Jin et al., 2025]. Our headline metric is the *misalignment score* $\text{MIS}(c) = \|\hat{m}^{(c)} - h^{(c)}\|_2$, the straight-line distance between the model's preference vector and the country's human one; lower is better. We macro-average it over a 20-country panel that spans four continents. Every scenario is presented in the country's native language, and only the decision tokens A and B stay in English so that token extraction is consistent across model families. Pearson r (shape agreement) and the base-2 Jensen-Shannon distance (JSD) serve as secondary diagnostics. The country list, the preprocessing (deduplication, quality filter, per-attribute caps, oversampling and deterministic shuffling), the per-country slice sizes and a cap-disabled sensitivity rerun are in App. A16.

We screened 22 open-weight models from 270M to 70B parameters on a five-country prototyping panel (App. A7). We report the seven backbones we evaluated on the full 20-country grid: Llama-3.3-70B, Magistral-24B, Phi-4 (14B), Qwen3-VL-8B, Qwen2.5-7B, Phi-3.5-mini (3.8B), and Gemma-4-E2B (2B). Llama-3.3-70B and Qwen3-VL-8B run from 4-bit quantised checkpoints and Qwen2.5-7B in BF16. For the other four the precision was not logged, and the loader's default was 4-bit, so comparisons across sizes are also comparisons across precision. The hyperparameters were checked on a small prototyping panel of three backbones and five countries and then frozen for the headline sweep. App. A15 describes how we elicit A and B consistently across model families.

We compare DISCA with training-free baselines that never see the human AMCE, all on the same 20-country Phi-4 grid (Table 1; App. A12): vanilla decoding, a WVS profile prompt that summarises the country's WVS values, a persona-ensemble prompt that applies pluralistic persona prompting [Santurkar et al., 2023, Sorensen et al., 2024], a fixed logit offset, activation steering [Arditi et al., 2024, Zou et al., 2023], and MC-Dropout [Gal and Ghahramani, 2016, Kwon et al., 2026]. We exclude reward-guided decoding [Khanov et al., 2024, Mudgal et al., 2024], because the per-country reward models it assumes do not exist for most countries.

4.1 Main results

A single decoding-time intervention, with no weight changes and no per-country labels, closes much of the gap between LLMs and human moral preferences. DISCA reduces macro MIS by 10 to 24%

Table 1: Training-free baselines on the Phi-4 20-country grid (macro MIS, lower is better). No method receives the human AMCE or updates weights.

Method	MIS ↓
DISCA (ours)	**0.346**
Persona-ensemble prompt	0.364
MC-Dropout [Gal and Ghahramani, 2016]	0.403
Activation steering [Arditi et al., 2024]	0.430
Fixed logit offset	0.439
WVS Profile Prompt	0.452
Vanilla decoding	0.454

Table 2: Macro MIS on the 20-country MultiTP slice, one run per backbone (seed 42; per country in App. A4). Gain: relative change of the macro MIS means. Win: countries where DISCA has lower MIS than vanilla. r: mean per-country Pearson correlation with the human AMCE.

Model	MIS ↓	Gain (%)	Win/20	r
Llama-3.3-70B	.668	+21.4	**20**	−0.01
Magistral-24B	.350	+13.1	16	+0.62
Phi-4 (14B)	**.346**	**+23.7**	18	+0.56
Qwen3-VL-8B	.466	+18.9	**20**	+0.22
Qwen2.5-7B	.362	+20.1	17	+0.45
Phi-3.5-mini	.571	+10.4	17	−0.35
Gemma-4-E2B (2B)	.478	+3.4	14	+0.46

Table 3: What each stage of DISCA contributes (macro MIS, lower is better), replayed from the logged controller state of the seven headline runs (seed 42). Arms are defined in §4.2. Bold marks the lowest value per backbone. Country win counts are in App. A10.

Arm	Llama-70B	Magistral	Phi-4	Qwen3-VL	Qwen2.5-7B	Phi-3.5	Gemma-E2B
Vanilla	.849	.403	.454	.575	.453	.638	.495
Consensus	.794	.330	.378	.564	.373	.649	.508
Consensus + prior	**.643**	.350	.344	.465	**.338**	**.562**	.480
DISCA without prior	.830	**.302**	.359	.549	.389	.668	.498
DISCA without refinement	.665	.349	**.344**	**.464**	.359	.570	**.474**
Full DISCA	.668	.350	.346	.466	.362	.571	.478

relative to vanilla decoding on the six backbones of 3.8B parameters or more (Table 2). Llama-3.3-70B and Qwen3-VL-8B improve on every one of the twenty countries. Figure 3 (App. A6) shows the same effect geometrically for Llama-3.3-70B: every country point moves from the vanilla cloud toward the human cloud. On the same Phi-4 grid, DISCA has the lowest MIS of the training-free baselines (Table 1).

The gains are spread across countries rather than concentrated in a few lucky cells (Table 7, App. A4). Eight robustness checks are collected in App. A13, and a three-seed rerun on two Qwen2.5 backbones supplies the controls of §4.3 (App. A9.1).

A smaller model can end better aligned. A 14B model with DISCA (Phi-4) reaches lower absolute misalignment than a 70B model without it (Llama-3.3-70B; vanilla values in Table 7). This compares two backbones, not a scaling law, since vanilla Phi-4 already starts much closer. The per-dimension error analysis (Table 8, App. A5) explains why: different models hit a wall on different attributes. Weaker models are dominated by Utilitarianism errors that are too large for a persona-based correction to close. Better-calibrated models have already resolved Utilitarianism and Species, so their bottleneck moves to Social Value, which DISCA improves, and Age, which it does not. Phi-4 is the most balanced of the seven backbones across all dimensions, and this balance, more than parameter count, explains its lead.

4.2 What each component contributes

The persona panel and the country prior together carry the gain (Table 3). We measure each stage by replaying, on CPU, the logged per-scenario controller state of the seven headline runs with that stage switched off. No model is re-run, and the Full DISCA row reproduces Table 2 exactly. *Consensus* decodes the mean persona gap $\bar{\delta}$ directly. *Consensus + prior* passes $\bar{\delta}$ in place of δ_{micro} through the running prior of Eq. 8. *DISCA without prior* emits δ_{micro} of Eq. 7, so $a_t{=}0$ throughout. *DISCA without refinement* sets $\delta^{\star}{=}0$ and keeps the ESS anchor and the prior.

The panel is the first source of gain: its plain consensus already lowers MIS on five of the seven backbones. The country prior is the second. The panel lowers MIS more than the prior does on Magistral, Phi-4 and Qwen2.5-7B, and the prior does more on Llama-3.3-70B, Qwen3-VL-8B, Phi-3.5-mini and Gemma-4-E2B. With both in place, Full DISCA is below vanilla decoding on every

Table 4: Same-logit controls and transfer to NormAd on Qwen2.5-7B / 14B. MultiTP rows reuse the saved logits of a separate set of three-seed runs (mean MIS ↓). These separate decodes give vanilla and DISCA values that differ from the headline sweep of Table 2 (Qwen2.5-7B: .453 and .362 there). Aggregation rules are defined in App. A11. NormAd: equal-country accuracy ↑, country-context condition, 17 countries, one run (seed 42); n/a = not computed. Full tables, including GlobalOpinionQA, are in App. A11.

Method	MultiTP MIS ↓	NormAd Acc. ↑
Vanilla	.423 / .403	.554 / .530
Probability averaging	.387 / .434	.576 / .550
Logit consensus	.388 / .439	n/a
Majority vote	.431 / .461	.573 / .546
DISCA	**.356 / .395**	**.577 / .553**

backbone. Removing the gated refinement lowers macro MIS slightly on every backbone, by less than 0.005. Its role is to keep each refinement bounded: on any scenario it can move the score by no more than the largest perturbation drawn (Theorem 2). Further comparisons between the arms, and an oracle calibration control that uses the human labels, are reported in App. A10.

4.3 Controls, transfer and scope

DISCA beats the standard ways of aggregating the panel on exactly the same logits (Table 4). The comparison reuses the saved base and persona gaps of three-seed runs on Qwen2.5-7B and 14B, so no comparator calls the model again and the only difference is how the panel is aggregated (App. A11). The standard rules run without the country prior, so the comparison credits the full controller, not a single stage.

The panel signal transfers beyond crash dilemmas. On NormAd [Rao et al., 2025], a three-way social-norm benchmark, DISCA raises accuracy by about two points over vanilla on both backbones when the prompt gives only the country, and edges out probability averaging over the personas (Table 4; the richer prompt conditions are in App. A11). On GlobalOpinionQA [Durmus et al., 2024], DISCA moves answers closer to the survey distributions than vanilla decoding; plain probability averaging over the same personas is closer still (App. A11).

The controller responds to what the personas say, not to how they are listed. Permuting the order of the four personas leaves every decision unchanged on a fresh 310-scenario United States audit with Qwen2.5-7B: across canonical, reversed and seeded random orders, no final sparing probability moves and no hard decision changes (App. A11). The content of the profile also has an effect. A wrong-country test keeps the scenarios, the prompt language and the country named in each persona header, and replaces only the ten-dimensional WVS profile behind the four personas with another country's profile, either a random one or the most distant one. This swap raises macro MIS in all four conditions on the two Qwen2.5 backbones, by up to about 5% on Qwen2.5-14B and by under 1% on Qwen2.5-7B, and the matched profile is better in only 10 to 12 of the 20 countries (App. A14.5).

Failures come from the backbone's logits, not from the culture. In the 22-model screening sweep (App. A7), success tracks how well the decision logits are conditioned, not parameter count: Phi-4 leads at 14B while several 30B+ models degrade. Statistics read from the vanilla pass, such as decision margin and entropy, are natural candidates for a deployment-time check that skips poorly conditioned models. The price is 3.57 to 3.72 times vanilla latency (App. A17).

The signal stops at factual recall. On BLEnD factual cultural QA [Myung et al., 2024] it does not help (App. A8). MultiTP value alignment lives in a scalar logit gap that importance sampling can search well. Factual QA picks one token from a vocabulary of about 100k entries for Phi-4, where perturbations tuned for a binary gap become noise on a wide softmax.

5 Limitations and Ethical Considerations

Personas are statistical aggregates, not people. Each persona summarises World Values Survey cohort statistics, which is what lets a public survey replace per-country labels. A persona describes no individual, and presenting it as one would flatten identity and risk stereotyping.

Country is a coarse proxy for culture. National aggregation hides sub-national, diasporic, indigenous, migrant, political and religious plurality, and survey coverage is fixed in time and uneven across countries. The panel needs only a description of a population, so the same controller can serve regional or community panels once such surveys exist.

Alignment is not legitimacy. MIS, accuracy and JSD measure agreement with a descriptive target [Awad et al., 2018, Jin et al., 2025]. They do not show that a majority preference ought to govern a decision, a question that needs normative and participatory input beyond any benchmark. Survey-based alignment scores are themselves unstable [Khan et al., 2025].

Minority risk remains. In 51.4% and 53.3% of scenarios on Qwen2.5-7B and 14B (three-seed means, App. A9.1), the correction moves the decision further from at least one synthetic persona's own gap than vanilla decoding was. The optional per-persona utility floor (App. A13, Table 27) and the loss-averse kernel are heuristics, not a fairness guarantee, and an inference-time controller deployed naively can encode a harmful majority view. The panel exposes every persona's utility, so this risk can be monitored per scenario, and turning that signal into a formal minority guarantee is open work.

No consequential individual use. DISCA should not drive decisions about individuals in healthcare, education, hiring, moderation sanctions, religion or family matters, and it must never infer a person's values from their nationality.

Method scope. The country prior is transductive, so each decision depends on earlier scenarios of the same country and on their order. Estimating the prior from public data before deployment would remove this dependence. The prior contributes more than the persona panel on four of the seven backbones (Table 3), so a prior that can be deployed is the main open step. DISCA needs decision-token scores, which text-only APIs do not always expose, and the Prospect-Theory function is an aggregation kernel, not a cognitive model. Each country is tied to one target language, so separating language from country effects needs a design that crosses the two. App. A19 lists further limitations and the claims this evidence does not support.

6 Conclusion

LLMs need not be retrained to respect moral diversity. We introduced *disagreement-driven steering*, which uses the disagreement among grounded persona agents to decide how much of a correction to trust. DISCA applies this idea to cultural alignment. Without changing a single weight, it reduces misalignment by 10 to 24% on the six backbones of 3.8B parameters or more, across twenty countries. With DISCA, a 14B model is better aligned than a vanilla 70B model. The mechanism is concrete. A panel of survey-grounded personas supplies the signal, a running prior pools decisions within a country, and a gated refinement keeps the correction bounded. A component decomposition shows that the panel and the prior together carry the gain. On MultiTP, DISCA outperforms every training-free baseline we test and, on the same logits, the standard aggregation rules applied without the country prior. The same panel signal improves social-norm classification on NormAd when the prompt gives only the country, and stops at a clear boundary on factual recall. Two directions follow. The transductive country prior can be replaced by one estimated from public data before deployment, and the panel signal can be tested on tasks where individuals, not countries, are the unit of alignment.

Acknowledgements and Disclosure of Funding

We thank the anonymous reviewers and the area chair for their constructive feedback, which improved this paper. Experiments were run on the free GPU quota provided by Kaggle and on a single H100 GPU. This work received no other third-party funding, and the authors declare no competing interests.

References

A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee, and N. Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, pages 136037–136083, 2024. doi: 10.52202/079017-4322.

M. Atari, M. J. Xue, P. S. Park, D. E. Blasi, and J. Henrich. Which humans? *PsyArXiv preprint*, 2023. doi: 10.31234/osf.io/5b26t. https://osf.io/preprints/psyarxiv/5b26t.

E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. The Moral Machine experiment. *Nature*, 563(7729):59–64, 2018. doi: 10.1038/s41586-018-0637-6.

M. S. Z. bin Ahmad and K. Takemoto. Large-scale moral machine experiment on large language models. *PLOS ONE*, 20(5):e0322776, 2025. doi: 10.1371/journal.pone.0322776. URL https://doi.org/10.1371/journal.pone.0322776.

R. Chen, W. Chai, Z. Yang, X. Zhang, Z. Wang, T. Quek, J. T. Zhou, S. Poria, and Z. Liu. DiffPO: Diffusion-styled preference optimization for inference time alignment of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18910–18925, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.926. URL https://aclanthology.org/2025.acl-long.926/. arXiv:2503.04240.

Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html. arXiv:2305.14325.

E. Durmus, K. Nguyen, T. I. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, Z. Hatfield-Dodds, D. Hernandez, N. Joseph, L. Lovitt, S. McCandlish, O. Sikder, A. Tamkin, J. Thamkul, J. Kaplan, J. Clark, and D. Ganguli. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling (COLM)*, 2024. arXiv:2306.16388.

K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 12634–12651. PMLR, 2024. URL https://proceedings.mlr.press/v235/ethayarajh24a.html. arXiv:2402.01306 (as "KTO: Model Alignment as Prospect Theoretic Optimization").

Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059. PMLR, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

C. M. Greco, L. La Cava, and A. Tagarelli. Culturally grounded personas in large language models: Characterization and alignment with socio-psychological value frameworks. *arXiv preprint arXiv:2601.22396*, 2026. doi: 10.48550/arXiv.2601.22396.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html. arXiv:1706.04599.

C. Haerpfer, R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, and B. Puranen. World Values Survey Wave 7 (2017-2022) Cross-National Data-Set, Version 6.0. World Values Survey Association, 2024. doi: 10.14281/18241.24.

J. Hainmueller, D. J. Hopkins, and T. Yamamoto. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, 22(1): 1–30, 2014. doi: 10.1093/pan/mpt024.

J. Henrich, S. J. Heine, and A. Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83, 2010. doi: 10.1017/S0140525X0999152X.

R. Inglehart and C. Welzel. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge University Press, Cambridge, 2005. doi: 10.1017/CBO9780511790881.

Z. Jin, M. Kleiman-Weiner, G. Piatti, S. Levine, J. Liu, F. Gonzalez, F. Ortu, A. Strausz, M. Sachan, R. Mihalcea, Y. Choi, and B. Schölkopf. Language model alignment in multilingual trolley problems. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VEqPDZIDAh. arXiv:2407.02273.

D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979. doi: 10.2307/1914185.

E. Kalai and M. Smorodinsky. Other solutions to Nash's bargaining problem. *Econometrica*, 43(3): 513–518, 1975. doi: 10.2307/1914280.

A. Khan, S. Casper, and D. Hadfield-Menell. Randomness, not representation: The unreliability of evaluating cultural alignment in LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2151–2165. Association for Computing Machinery, 2025. doi: 10.1145/3715275.3732147. URL https://dl.acm.org/doi/10.1145/3715275.3732147. arXiv:2503.08688.

M. Khanov, J. Burapacheep, and Y. Li. ARGS: Alignment as reward-guided search. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shgx0eqdw6. arXiv:2402.01694.

J. Kim, J. Kwon, L. F. Vecchietti, A. Oh, and M. Cha. Exploring persona-dependent LLM alignment for the Moral Machine experiment. *arXiv preprint arXiv:2504.10886*, 2025. doi: 10.48550/arXiv.2504.10886. URL https://arxiv.org/abs/2504.10886.

J. Kwon, L. F. Vecchietti, S. Park, and M. Cha. Dropouts in confidence: Moral uncertainty in human-LLM alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37547–37555, 2026. doi: 10.1609/aaai.v40i44.41088. arXiv:2511.13290.

S. Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018. doi: 10.48550/arXiv.1805.00909. URL https://arxiv.org/abs/1805.00909.

T. Liang, Z. He, W. Jiao, X. Wang, Y. Wang, R. Wang, Y. Yang, S. Shi, and Z. Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

P. C. Mahalanobis. Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109(4):325–378, 1946.

P. J. McCarthy. Pseudo-replication: Half samples. *Review of the International Statistical Institute*, 37(3):239–264, 1969. doi: 10.2307/1402116.

S. Mudgal, J. Lee, H. Ganapathy, Y. Li, T. Wang, Y. Huang, Z. Chen, H.-T. Cheng, M. Collins, T. Strohman, J. Chen, A. Beutel, and A. Beirami. Controlled decoding from language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36486–36503. PMLR, 2024. URL https://proceedings.mlr.press/v235/mudgal24a.html. arXiv:2310.17022.

J. Myung, N. Lee, Y. Zhou, J. Jin, R. A. Putri, D. Antypas, H. Borkakoty, E. Kim, C. Perez-Almendros, A. A. Ayele, V. Gutiérrez-Basulto, Y. Ibáñez-García, H. Lee, S. H. Muhammad, K. Park, A. S. Rzayev, N. White, S. M. Yimam, M. T. Pilehvar, N. Ousidhoum, J. Camacho-Collados, and A. Oh. BLEnD: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 37, pages 78104–78146, 2024. doi: 10.52202/079017-2483.

J. F. Nash. The bargaining problem. *Econometrica*, 18(2):155–162, 1950. doi: 10.2307/1907266.

A. Rao, A. Yerukola, V. Shah, K. Reinecke, and M. Sap. NormAd: A framework for measuring the cultural adaptability of 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 2373–2403. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.120.

M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. *Electronic Communications in Probability*, 18(82):1–9, 2013. doi: 10.1214/ECP.v18-2865.

M. J. Ryan, W. Held, and D. Yang. Unintended impacts of LLM alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16121–16140, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.853. URL https://aclanthology.org/2024.acl-long.853/. arXiv:2402.15018.

S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. 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, 2023. URL https://proceedings.mlr.press/v202/santurkar23a.html.

T. Sorensen, J. Moore, J. Fisher, M. L. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, T. Althoff, and Y. Choi. 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, 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

K. Takemoto. The moral machine experiment on large language models. *Royal Society Open Science*, 11(2):231393, 2024. doi: 10.1098/rsos.231393.

A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023. doi: 10.48550/arXiv.2308.10248.

A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992. doi: 10.1007/BF00122574. URL https://link.springer.com/article/10.1007/BF00122574.

X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw. arXiv:2203.11171.

G. Williams, P. Drews, B. Goldfain, J. M. Rehg, and E. A. Theodorou. Information-theoretic model predictive control: Theory and applications to autonomous driving. *IEEE Transactions on Robotics*, 34(6):1603–1622, 2018. doi: 10.1109/TRO.2018.2865891. URL https://ieeexplore.ieee.org/abstract/document/8558663.

K. M. Wolter. *Introduction to Variance Estimation*. Springer, New York, 2nd edition, 2007. doi: 10.1007/978-0-387-35099-8.

J. Yao, X. Yi, J. Wang, Z. Dou, and X. Xie. CAReDiO: Cultural alignment via representativeness and distinctiveness guided data optimization. *arXiv preprint arXiv:2504.08820*, 2025. doi: 10.48550/arXiv.2504.08820.

A. Zewail, A. Figueroa, J. Graham, and M. Atari. Moral stereotyping in large language models. *Proceedings of the National Academy of Sciences*, 123(10):e2519941123, 2026. doi: 10.1073/pnas.2519941123. URL https://www.pnas.org/doi/10.1073/pnas.2519941123.

B. Zhang, X. Zhao, J. Li, H. Chen, and Z. Chen. Mind the gap in cultural alignment: Task-aware culture management for large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16798–16813. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.766.

A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. doi: 10.48550/arXiv.2310.01405.

Appendix Contents

A1 MultiTP Dimensions and DISCA Inference Pseudocode **17**

A2 Proof of Proposition 1 **17**

A3 Additional Theoretical Properties of DISCA **20**

A3.1 Bounded refinement 20

A3.2 Finite-sample concentration of the within-panel variance estimator 22

A3.3 Hölder stability with respect to backbone perturbations 24

A4 Full Per-Country Results **27**

A5 Per-Dimension Error Analysis **27**

A5.1 Diagnosing negative Pearson r 29

A6 Geometric and Per-Country Visualizations **30**

A7 Broader Model Landscape **30**

A8 Factual Cultural QA Evaluation (BLEnD) **31**

A9 Post-Hoc Diagnostics **32**

A9.1 Seeds 32

A9.2 Scenario-level conditioning of the DISCA correction 32

A10 Component Decomposition **33**

A11 Same-Logit Controls and Transfer Benchmarks **34**

A12 Baseline Implementation Details **36**

A12.1 Other training-free baselines (Table 1) 36

A12.2 Activation Steering 37

A12.3 ARGS and controlled decoding (relation to our work) 37

A12.4 Additional inference-time baselines (twenty-country Phi-4 grid) 37

A13 Hyperparameters and Sensitivity Sweeps **38**

A13.1 Robustness Summary 38

A13.2 Default Hyperparameters and Validation 39

A13.3 Temperature Sensitivity 39

A13.4 Extended Hyperparameter Sensitivity 40

A14 Persona Construction Details **42**

A14.1 Persona Prompts 42

A14.2 WVS Data Processing Pipeline 44

A14.3 WVS-to-Trolley Dimension Linkage 45

A14.4 Macro-level WVS dimension importance 46

A14.5 Wrong-country profile test 47

A15 AMCE Estimation Details **47**

A16 Dataset Preprocessing and Sensitivity **48**

A17 Trigger Mechanism and Latency **49**

A18 Relationship to Persona-Dependent LLM Alignment **50**

A19 Extended Limitations **51**

A1 MultiTP Dimensions and DISCA Inference Pseudocode

The pipeline schematic is Figure 1 in §3. Table 5 lists the six MultiTP moral dimensions and the per-attribute logit temperatures T_c used by DISCA (validated in App. A13). Algorithm 1 states the exact inference routine.

Algorithm 1 DISCA Inference: Scenario $\mathbf{x}$ in Country c

Require: Frozen LLM f_θ; personas $\{\mathbf{s}_i\}_{i=1}^N$; hyperparameters $K_{\text{half}}, \sigma, \lambda, \eta, T_{\text{dec}}, s$
Ensure: $p_{\text{spare}} \in [0, 1]$

for order $\in \{\text{orig}, \text{swap}\}$ **do**
  Batched forward: f_θ on base + N personas
  Extract logit pairs; compute $\delta_i^{(\text{ord})}$
end for
$\delta_i \leftarrow (\delta_i^{(\text{orig})} - \delta_i^{(\text{swap})})/2$
Compute consensus $\bar{\delta} \leftarrow \frac{1}{N}\sum_i \delta_i$
for $m \in \{1, 2\}$ **do**
  for $k = 1$ **to** K_{half} **do**
    Sample $\epsilon_k \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = \max\{\text{sd}(\delta_{1:N}), 0.3\}$; compute $U_{\text{total}}(\epsilon_k)$ via Eqs. 1, 3
  end for
  $\delta_m^\star \leftarrow \sum_k w_k^{(m)} \epsilon_k$ if $\text{ESS}_m \geq \rho_{\text{eff}}$ else 0
end for
$r \leftarrow \exp(-(\delta_1^\star - \delta_2^\star)^2/s)$
$\delta^\star \leftarrow r\,(\delta_1^\star + \delta_2^\star)/2$
$\delta_{\text{micro}} \leftarrow \alpha_{\text{ess}}\,\bar{\delta} + (1 - \alpha_{\text{ess}})\,\delta_{\text{base}} + \delta^\star$ (with $\alpha_{\text{ess}} = \min\{1, \max(\rho_{\text{eff}}, \min(\text{ESS}_1, \text{ESS}_2))\}$, $\text{ESS}_m = (\sum_k w_k^{(m)2})^{-1}/K_{\text{half}}$)
$a_t \leftarrow 0$ if $t < t_0$ else $1 - e^{-(t-t_0)/\tau}$; $\delta_{\text{final}} \leftarrow a_t\,d_c + (1 - a_t)\,\delta_{\text{micro}}$ (country prior, §3.3)
$d_c \leftarrow (1 - \omega)\,d_c + \omega\,\delta_{\text{micro}}$; $t \leftarrow t + 1$
return $p_{\text{spare}} = \sigma(\delta_{\text{final}}/T_{\text{dec}})$

Table 5: The six moral dimensions in MultiTP. T_c is the per-attribute logit temperature used by our method; values reflect empirical logit magnitudes and are validated in App. A13.

Dimension	Preferred	Contra	T_c
Species	Human	Animal	4.0
Gender	Female	Male	3.5
Age	Young	Elderly	1.5
Fitness	Fit	Unfit	1.5
Social Value	High (exec.)	Low (homeless)	1.5
Utilitarianism	More lives	Fewer lives	1.5

A2 Proof of Proposition 1

Recall the agent model from §3.2: for a single scenario,

$$\delta_i = \delta_h + \eta_i, \qquad i = 1, \ldots, N,$$

where δ_h is a fixed (but unknown) population-preference logit gap, and the noise terms η_i are i.i.d. with $\mathbb{E}[\eta_i] = 0$ and $\text{Var}(\eta_i) = \tau^2$. We also fix δ_{base}, the gap from the base prompt (deterministic given the scenario), and define

$$\bar{\delta} = \frac{1}{N}\sum_{i=1}^N \delta_i, \quad D^2 = \frac{1}{N-1}\sum_{i=1}^N (\delta_i - \bar{\delta})^2, \quad \Delta = \bar{\delta} - \delta_{\text{base}}, \quad \Delta_h = \delta_h - \delta_{\text{base}}.$$

Throughout, expectations and variances are taken over the noise $\eta_1, \ldots, \eta_N$.

Part (i): $\mathbb{E}[D^2] = \tau^2$ and $\mathrm{Var}(\Delta) = \tau^2/N$

We first rewrite D^2 in terms of the noise variables. For each i,

$$\delta_i - \bar{\delta} = (\delta_h + \eta_i) - \Big(\delta_h + \tfrac{1}{N}\sum_j \eta_j\Big) = \eta_i - \bar{\eta},$$

where $\bar{\eta} = \frac{1}{N}\sum_j \eta_j$. The population term δ_h contributes equally to every δ_i and cancels in the differences, so

$$D^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\eta_i - \bar{\eta})^2. \tag{10}$$

Expanding the square and using $\sum_i \eta_i = N\bar{\eta}$,

$$\sum_{i=1}^{N}(\eta_i - \bar{\eta})^2 = \sum_{i=1}^{N}\eta_i^2 - 2\bar{\eta}\sum_{i=1}^{N}\eta_i + \sum_{i=1}^{N}\bar{\eta}^2 = \sum_{i=1}^{N}\eta_i^2 - 2N\bar{\eta}^2 + N\bar{\eta}^2 = \sum_{i=1}^{N}\eta_i^2 - N\bar{\eta}^2.$$

Taking expectations on both sides,

$$\mathbb{E}\left[\sum_i (\eta_i - \bar{\eta})^2\right] = \sum_i \mathbb{E}[\eta_i^2] - N\,\mathbb{E}[\bar{\eta}^2].$$

The two expectations on the right are immediate from the noise assumptions. For each i, $\mathbb{E}[\eta_i^2] = \mathrm{Var}(\eta_i) + (\mathbb{E}[\eta_i])^2 = \tau^2$, so $\sum_i \mathbb{E}[\eta_i^2] = N\tau^2$. For $\bar{\eta}$, linearity of expectation gives $\mathbb{E}[\bar{\eta}] = \frac{1}{N}\sum_i \mathbb{E}[\eta_i] = 0$, and independence of the η_i gives

$$\mathrm{Var}(\bar{\eta}) = \frac{1}{N^2}\sum_i \mathrm{Var}(\eta_i) = \frac{1}{N^2}\cdot N\tau^2 = \frac{\tau^2}{N},$$

so $\mathbb{E}[\bar{\eta}^2] = \mathrm{Var}(\bar{\eta}) + (\mathbb{E}[\bar{\eta}])^2 = \tau^2/N$. Substituting,

$$\mathbb{E}\left[\sum_i (\eta_i - \bar{\eta})^2\right] = N\tau^2 - N\cdot\frac{\tau^2}{N} = (N-1)\tau^2,$$

and dividing by $N-1$ gives

$$\mathbb{E}[D^2] = \tau^2. \tag{11}$$

The variance of the consensus adjustment follows directly. Since δ_{base} is deterministic,

$$\mathrm{Var}(\Delta) = \mathrm{Var}(\bar{\delta} - \delta_{\text{base}}) = \mathrm{Var}(\delta_h + \bar{\eta}) = \mathrm{Var}(\bar{\eta}) = \frac{\tau^2}{N}.$$

The two quantities just established play distinct roles: D^2 has expected value τ^2, the variance of one agent's noise (this is what we estimate from data), whereas $\mathrm{Var}(\Delta) = \tau^2/N$ is the variance of the mean of N agents' noise (this is what enters $\gamma^\star$ in Part (ii)).

Part (ii): MSE-minimising shrinkage

We seek the value of $\gamma \in [0,1]$ that minimises

$$\mathrm{MSE}(\gamma) = \mathbb{E}\Big[\big(\gamma\Delta - \Delta_h\big)^2\Big].$$

The strategy is to rewrite the error $\gamma\Delta - \Delta_h$ as a sum of a mean-zero random part and a deterministic constant, so that the cross term in the squared expansion vanishes. From Part (i), $\Delta = \delta_h + \bar{\eta} - \delta_{\text{base}}$ and $\mathbb{E}[\bar{\eta}] = 0$, so

$$\mathbb{E}[\Delta] = \delta_h + 0 - \delta_{\text{base}} = \Delta_h.$$

Adding and subtracting $\gamma\Delta_h$ inside the error,

$$\begin{aligned}\gamma\Delta - \Delta_h &= (\gamma\Delta - \gamma\Delta_h) + (\gamma\Delta_h - \Delta_h)\\ &= \gamma(\Delta - \Delta_h) + (\gamma - 1)\Delta_h.\end{aligned}$$

Squaring and taking expectations,

$$\begin{aligned}\mathrm{MSE}(\gamma) &= \gamma^2\,\mathbb{E}[(\Delta-\Delta_h)^2] + 2\gamma(\gamma-1)\Delta_h \underbrace{\mathbb{E}[\Delta-\Delta_h]}_{=0} + (\gamma-1)^2\Delta_h^2 \\ &= \gamma^2\cdot\frac{\tau^2}{N} + (\gamma-1)^2\Delta_h^2,\end{aligned}$$

where $\mathbb{E}[(\Delta-\Delta_h)^2] = \mathrm{Var}(\Delta) = \tau^2/N$ from Part (i). The two surviving terms are the standard bias-variance pieces:

- $\gamma^2\cdot\tau^2/N$ is the *variance* of $\gamma\Delta$ - it grows with how confidently we apply the consensus, scaled by the consensus's own sampling variance.
- $(\gamma-1)^2\Delta_h^2$ is the *squared bias* from shrinkage - zero at $\gamma=1$ (full adjustment, no bias), quadratic as $\gamma\to 0$.

Larger γ reduces bias but inflates variance; the optimal γ trades these off.

To find $\gamma^\star$, differentiate $\mathrm{MSE}(\gamma)$:

$$\begin{aligned}\frac{d}{d\gamma}\mathrm{MSE}(\gamma) &= 2\gamma\cdot\frac{\tau^2}{N} + 2(\gamma-1)\Delta_h^2 \\ &= 2\left[\gamma\left(\frac{\tau^2}{N}+\Delta_h^2\right) - \Delta_h^2\right].\end{aligned}$$

Setting this to zero gives

$$\gamma^\star = \frac{\Delta_h^2}{\Delta_h^2+\tau^2/N}. \tag{12}$$

The second derivative is $2(\Delta_h^2+\tau^2/N) > 0$ whenever $\tau^2>0$, so $\gamma^\star$ is the unique minimum.

It remains to check that $\gamma^\star$ behaves as the proposition claims. The numerator is non-negative and the denominator is at least as large, so $\gamma^\star\in[0,1]$ automatically. Writing $\gamma^\star$ as a function of $u := \tau^2/N$,

$$\gamma^\star(u) = \frac{\Delta_h^2}{\Delta_h^2+u}, \qquad \frac{d\gamma^\star}{du} = -\frac{\Delta_h^2}{(\Delta_h^2+u)^2} \le 0,$$

so $\gamma^\star$ is monotone-decreasing in the consensus variance, recovering the two boundary regimes:

- $\tau^2\to 0$ (agents agree): $\gamma^\star\to 1$, apply the consensus in full.
- $\tau^2\to\infty$ or $\Delta_h\to 0$ (signal-to-noise collapses): $\gamma^\star\to 0$, do not adjust at all.

□

Corollary 1: Marginal Value of Additional Personas

Corollary 1 (Marginal value of additional personas)**.** *Under the agent model of Proposition 1, the marginal reduction in consensus variance from adding the* $(N{+}1)$*-th persona is*

$$\Delta\mathrm{Var}(N\to N{+}1) \;=\; \frac{\tau^2}{N} - \frac{\tau^2}{N+1} \;=\; \frac{\tau^2}{N(N+1)} \;=\; \Theta(N^{-2}). \tag{13}$$

Equivalently, the optimal shrinkage coefficient $\gamma^\star(N) = \Delta_h^2/(\Delta_h^2+\tau^2/N)$ *from Proposition 1(ii) satisfies* $\gamma^\star(N{+}1)-\gamma^\star(N) = \Theta(N^{-2})$ *for* $N \gg \tau^2/\Delta_h^2$.

Proof. Direct calculation from $\mathrm{Var}(\Delta)=\tau^2/N$ (Proposition 1(i)). The $\gamma^\star$ statement follows from

$$\gamma^\star(N{+}1)-\gamma^\star(N) = \frac{\Delta_h^2\,\tau^2}{N(N+1)\,(\Delta_h^2+\tau^2/N)\,(\Delta_h^2+\tau^2/(N{+}1))},$$

which decays as $\Theta(N^{-2})$ in the regime $\tau^2/N \ll \Delta_h^2$. □ □

Expressing the marginal precision gain as a fraction of τ^2 exposes the diminishing-returns structure:

Table 6: Marginal reduction of $\mathrm{Var}(\Delta)$ from adding one persona, expressed as a fraction of the noise variance τ^2. The drop from $N{=}4$ to $N{=}5$ already sits below 5%, while compute scales linearly in N.

Transition	$\Delta\mathrm{Var}/\tau^2$	Fraction
$N{=}2 \to 3$	$1/6$	16.7%
$N{=}3 \to 4$	$1/12$	8.3%
$N{=}4 \to 5$	$1/20$	5.0%
$N{=}5 \to 6$	$1/30$	3.3%

Combined with the demographic-coverage requirement of $N \geq 3$ (three age cohorts) plus a country-wide aggregate as a fourth anchor, $N{=}4$ is the smallest panel that sits on the flat side of the $\Theta(N^{-2})$ diminishing-returns curve. Corollary 1 formalises this as a cost-precision inflection point rather than a global optimum: any panel with $N \geq 4$ achieves consensus variance within a factor 2 of the asymptotic floor, and Table 24 confirms the prediction empirically. We do not claim $N{=}4$ is uniquely best; we claim it is the smallest panel for which the marginal precision gain has decayed below the compute cost of an additional forward pass.

A3 Additional Theoretical Properties of DISCA

This appendix collects three structural properties of the DISCA pipeline that complement Proposition 1 and Corollary 1: a deterministic and high-probability bound on the correction magnitude (A3.1), finite-sample concentration of the within-panel variance estimator (A3.2), and stability of the correction under perturbations of the underlying LLM (A3.3).

Each result is stated for one scenario, given its persona gaps. The LLM enters only through those gaps and through the proposal scale $\sigma = \max\{\mathrm{sd}(\delta_{1:N}), 0.3\}$ they set; beyond that, the results depend only on hyperparameters of the controller, such as the softmax temperature η and the gate scale s. All of them concern the refinement $\delta^\star$, not the final decision gap, which also contains the anchor and the country prior.

A3.1 Bounded refinement

The dual-pass gate bounds the refinement term. We show that the gate combined with the IS aggregator yields a deterministic almost-sure bound on $|\delta^\star|$, plus a tighter high-probability bound that depends only on the proposal scale and the IS budget.

Theorem 2 (Bounded correction)**.** *Conditional on the realisation of the IS draws $\{\epsilon_k^{(m)}\}_{k,m}$, the DISCA correction satisfies*

$$|\delta^\star| \;\leq\; \max_{k=1,\ldots,K_{half};\; m\in\{1,2\}} \left|\epsilon_k^{(m)}\right| \tag{14}$$

almost surely. Marginally over the IS draws, for any failure probability $\delta_p \in (0,1)$,

$$|\delta^\star| \;\leq\; \sigma\sqrt{2\log(4K_{half}/\delta_p)} \tag{15}$$

holds with probability at least $1-\delta_p$. Here σ is the proposal scale of the scenario, $\sigma = \max\{\mathrm{sd}(\delta_{1:N}), 0.3\}$, so both bounds depend on the LLM and the scenario only through the spread of the persona gaps. The theorem bounds $\delta^\star$ only; the anchor and the country prior in δ_{final} are not covered.

Proof. We prove the deterministic bound (Eq. 14) first, then the high-probability bound (Eq. 15).

Deterministic bound. Each per-pass IS estimate $\delta^{(m)}_{\text{PT-IS}} = \sum_k w_k^{(m)} \epsilon_k^{(m)}$ is a convex combination of $\{\epsilon_k^{(m)}\}_{k=1}^{K_{\text{half}}}$, since the softmax weights $w_k^{(m)} \propto \exp(U_k^{(m)}/\eta)$ are non-negative and normalised to

sum to one. A convex combination of any set of values is bounded by the largest absolute value in that set, so

$$|\delta^{(m)}_{\text{PT-IS}}| \;\le\; \max_k |\epsilon^{(m)}_k| \qquad \text{for } m \in \{1,2\}.$$

The dual-pass average $(\delta^{(1)}_{\text{PT-IS}} + \delta^{(2)}_{\text{PT-IS}})/2$ is itself bounded by $\max_m |\delta^{(m)}_{\text{PT-IS}}|$, and the gate $r \in (0,1]$ from Eq. 5 only shrinks this further. Combining, $|\delta^\star| \le \max_{k,m} |\epsilon^{(m)}_k|$.

High-probability bound. For $X \sim \mathcal{N}(0,\sigma^2)$ and $t > 0$, the upper-tail probability is

$$\mathbb{P}(X \ge t) = \int_t^\infty \frac{1}{\sigma\sqrt{2\pi}}\, e^{-x^2/(2\sigma^2)}\, dx.$$

This integral has no closed form, but we can bound it by exploiting that $x \ge t$ throughout the domain of integration, so $1 \le x/t$. Multiplying the integrand by x/t only enlarges it:

$$\mathbb{P}(X \ge t) \;\le\; \frac{1}{t\sigma\sqrt{2\pi}} \int_t^\infty x\, e^{-x^2/(2\sigma^2)}\, dx.$$

The new integral has a closed form via the substitution $u = x^2/(2\sigma^2)$, $du = (x/\sigma^2)\, dx$:

$$\int_t^\infty x\, e^{-x^2/(2\sigma^2)}\, dx \;=\; \sigma^2 \int_{t^2/(2\sigma^2)}^\infty e^{-u}\, du \;=\; \sigma^2 e^{-t^2/(2\sigma^2)}.$$

Substituting back gives:

$$\mathbb{P}(X \ge t) \;\le\; \frac{\sigma}{t\sqrt{2\pi}}\, e^{-t^2/(2\sigma^2)}.$$

For $t \ge \sigma$, the prefactor $\sigma/(t\sqrt{2\pi}) < 1$, so we can drop it and absorb it into the looser-but-cleaner form

$$\mathbb{P}(X \ge t) \;\le\; \exp\!\left(-\frac{t^2}{2\sigma^2}\right).$$

By symmetry of the Gaussian about zero, $\mathbb{P}(X \le -t) = \mathbb{P}(X \ge t)$, so the two-sided bound is

$$\mathbb{P}(|\epsilon^{(m)}_k| \ge t) \;\le\; 2\exp\!\left(-\frac{t^2}{2\sigma^2}\right) \qquad \forall t \ge \sigma. \tag{16}$$

The event $\{\max_{k,m} |\epsilon^{(m)}_k| \ge t\}$ holds if and only if at least one individual draw exceeds t in absolute value. If some $|\epsilon^{(m)}_j| \ge t$, then the maximum, which is at least as large as every individual element, is also $\ge t$. So,

$$\mathbb{P}\left(\max_{k,m} |\epsilon^{(m)}_k| \ge t\right) \;=\; \mathbb{P}\left(\bigcup_{k,m} \{|\epsilon^{(m)}_k| \ge t\}\right)$$

The probability of a union is at most the sum of the individual probabilities (the union bound, also called Boole's inequality):

$$\mathbb{P}\left(\bigcup_{k,m} \{|\epsilon^{(m)}_k| \ge t\}\right) \;\le\; \sum_{k,m} \mathbb{P}(|\epsilon^{(m)}_k| \ge t).$$

Combining with the per-draw tail bound (Eq. 16):

$$\mathbb{P}\left(\max_{k,m} |\epsilon^{(m)}_k| \ge t\right) \;\le\; \sum_{k,m} \mathbb{P}(|\epsilon^{(m)}_k| \ge t) \;\le\; 4K_{\text{half}} \exp\!\left(-\frac{t^2}{2\sigma^2}\right),$$

where the factor $4K_{\text{half}} = 2 \cdot 2K_{\text{half}}$ comes from the two-sided constant in Eq. 16 times the $2K_{\text{half}}$ independent draws.

Setting the right-hand side equal to δ_p and solving for t:

$$4K_{\text{half}} \exp(-t^2/(2\sigma^2)) = \delta_p \iff t = \sigma\sqrt{2\log(4K_{\text{half}}/\delta_p)}.$$

Therefore $\max_{k,m} |\epsilon^{(m)}_k| \le \sigma\sqrt{2\log(4K_{\text{half}}/\delta_p)}$ with probability at least $1 - \delta_p$. Combining with the deterministic bound $|\delta^\star| \le \max_{k,m} |\epsilon^{(m)}_k|$ yields Eq. 15. □

With σ at its floor of 0.3, $K_{\text{half}} = 64$, and $\delta_p = 0.05$:

$$|\delta^\star| \;\le\; 0.3 \cdot \sqrt{2\log(5120)} \;\approx\; 0.3 \cdot \sqrt{17.07} \;\approx\; 1.24$$

in logit space with probability at least 0.95; the bound grows linearly when the persona spread raises σ above the floor. The gaps are already divided by the per-attribute temperatures, and p_{spare} is the sigmoid of $\delta_{\text{final}}/T_{\text{dec}}$ with $T_{\text{dec}} = 0.5$, so the maximum sigmoid slope is $1/(4T_{\text{dec}}) = 0.5$. At the floor, the refinement alone therefore moves p_{spare} by at most about 0.62. This is a loose worst-case bound, not a tight cap.

Scope of the bound. Theorem 2 bounds only the refinement $\delta^\star$. It says nothing about the anchor blend or the country prior, which act on the consensus and on earlier decisions.

A3.2 Finite-sample concentration of the within-panel variance estimator

Proposition 1(i) establishes that D^2 is unbiased for τ^2. Unbiasedness is the right starting point but not enough at $N{=}4$, where the deviation of D^2 from τ^2 on a single scenario can be substantial. We bound the deviation explicitly and use the result to motivate the dual-pass gate as a complementary variance estimator.

Theorem 3 (Concentration of D^2)**.** *Suppose the persona noise terms η_i in the agent model of Proposition 1 are independent sub-Gaussian random variables with parameter $\sigma_\eta \ge \tau$ (so that $\mathrm{Var}(\eta_i) = \tau^2 \le \sigma_\eta^2$). Then for any $t > 0$,*

$$\mathbb{P}\big(|D^2 - \tau^2| \ge t\big) \;\le\; 2\exp\left(-c(N-1)\min\left(\frac{t^2}{\sigma_\eta^4}, \frac{t}{\sigma_\eta^2}\right)\right) \tag{17}$$

for an absolute constant $c > 0$. Equivalently, with probability at least $1 - \delta$,

$$|D^2 - \tau^2| \;\le\; C\,\sigma_\eta^2\left(\sqrt{\frac{\log(2/\delta)}{N-1}} + \frac{\log(2/\delta)}{N-1}\right). \tag{18}$$

Proof. We rewrite D^2 as a quadratic form in $\boldsymbol{\eta}$, compute the relevant matrix norms, apply the Hanson-Wright inequality, and invert the resulting tail bound.

From the proof of Proposition 1(i), $\sum_i(\delta_i - \bar\delta)^2 = \sum_i(\eta_i - \bar\eta)^2$ (the δ_h terms cancel because they shift every entry equally). Define the centering matrix

$$H \;:=\; I - \frac{1}{N}\mathbf{1}\mathbf{1}^\top \;\in\; \mathbb{R}^{N\times N},$$

where $\mathbf{1}$ is the N-vector of ones. The action of H on a vector $\boldsymbol{\eta}$ is

$$H\boldsymbol{\eta} \;=\; \boldsymbol{\eta} - \frac{1}{N}\mathbf{1}\,(\mathbf{1}^\top\boldsymbol{\eta}) \;=\; \boldsymbol{\eta} - \bar\eta\,\mathbf{1},$$

so the i-th entry of $H\boldsymbol{\eta}$ is $\eta_i - \bar\eta$. The matrix H is symmetric, and a direct calculation using $\mathbf{1}^\top\mathbf{1} = N$ verifies idempotency:

$$H^2 \;=\; \big(I - \tfrac{1}{N}\mathbf{1}\mathbf{1}^\top\big)^2 \;=\; I - \tfrac{2}{N}\mathbf{1}\mathbf{1}^\top + \tfrac{1}{N^2}\mathbf{1}(\mathbf{1}^\top\mathbf{1})\mathbf{1}^\top \;=\; I - \tfrac{1}{N}\mathbf{1}\mathbf{1}^\top \;=\; H.$$

Therefore

$$\boldsymbol{\eta}^\top H\boldsymbol{\eta} \;=\; \boldsymbol{\eta}^\top H^\top H\boldsymbol{\eta} \;=\; (H\boldsymbol{\eta})^\top(H\boldsymbol{\eta}) \;=\; \sum_{i=1}^{N}(\eta_i - \bar\eta)^2.$$

Setting $A := H/(N-1)$ yields $D^2 = \boldsymbol{\eta}^\top A\boldsymbol{\eta}$, with $\mathbb{E}[D^2] = \tau^2$ from Proposition 1(i).

The centering matrix H is a rank $(N-1)$ orthogonal projection, with two eigenspaces:

- The all-ones direction $\mathbf{1}$ is an eigenvector with eigenvalue 0: $H\mathbf{1} = \mathbf{1} - \frac{1}{N}\mathbf{1}(\mathbf{1}^\top\mathbf{1}) = \mathbf{1} - \mathbf{1} = \mathbf{0}$.
- Any $\mathbf{v}$ orthogonal to $\mathbf{1}$ (i.e., $\mathbf{1}^\top\mathbf{v} = 0$) is an eigenvector with eigenvalue 1: $H\mathbf{v} = \mathbf{v} - \frac{1}{N}\mathbf{1}\cdot 0 = \mathbf{v}$.

The orthogonal complement of a single nonzero vector in $\mathbb{R}^N$ is $(N-1)$-dimensional, so H has eigenvalue 0 with multiplicity 1 and eigenvalue 1 with multiplicity $N-1$. Therefore $A = H/(N-1)$ has eigenvalues 0 (multiplicity 1) and $1/(N-1)$ (multiplicity $N-1$).

For a symmetric matrix, the operator norm is the largest absolute eigenvalue, and the squared Frobenius norm is the sum of squared eigenvalues (both follow from the spectral decomposition $A = Q\Lambda Q^\top$, since orthogonal similarity preserves traces and $\mathrm{tr}(A^2) = \mathrm{tr}(\Lambda^2) = \sum_i \lambda_i^2$). Therefore

$$\|A\|_{\mathrm{op}} \;=\; \frac{1}{N-1}, \qquad \|A\|_F^2 \;=\; \mathrm{tr}(A^2) \;=\; 0^2 \cdot 1 + \left(\frac{1}{N-1}\right)^2 (N-1) \;=\; \frac{1}{N-1}.$$

The Hanson-Wright inequality [Rudelson and Vershynin, 2013] states that for a vector $\boldsymbol{\eta}$ with independent mean-zero sub-Gaussian entries of parameter σ_η and any matrix A,

$$\mathbb{P}\big(|\boldsymbol{\eta}^\top A\boldsymbol{\eta} - \mathbb{E}[\boldsymbol{\eta}^\top A\boldsymbol{\eta}]| \geq t\big) \;\leq\; 2\exp\left(-c\min\left(\frac{t^2}{\sigma_\eta^4\|A\|_F^2},\, \frac{t}{\sigma_\eta^2\|A\|_{\mathrm{op}}}\right)\right) \tag{19}$$

for an absolute constant $c > 0$. Substituting $D^2 = \boldsymbol{\eta}^\top A\boldsymbol{\eta}$ with $\mathbb{E}[D^2] = \tau^2$, $\|A\|_F^2 = 1/(N-1)$, and $\|A\|_{\mathrm{op}} = 1/(N-1)$:

$$\mathbb{P}(|D^2 - \tau^2| \geq t) \;\leq\; 2\exp\left(-c(N-1)\min\left(\frac{t^2}{\sigma_\eta^4},\, \frac{t}{\sigma_\eta^2}\right)\right),$$

which is Eq. 17.

We seek the smallest threshold t at which the tail probability is at most δ. Setting the right-hand side of Eq. 17 less than or equal to δ, dividing by 2, taking logarithms, and multiplying by -1 (which flips $\log(\delta/2)$ to $\log(2/\delta)$):

$$c(N-1)\min\left(\frac{t^2}{\sigma_\eta^4},\, \frac{t}{\sigma_\eta^2}\right) \;\geq\; \log(2/\delta).$$

For the minimum of two non-negative quantities to be at least $\log(2/\delta)/(c(N-1))$, both arguments must individually satisfy this lower bound. This produces two regime-specific thresholds:

- $\dfrac{t^2}{\sigma_\eta^4} \geq \dfrac{\log(2/\delta)}{c(N-1)}$ gives $t \geq t_1 := \sigma_\eta^2\sqrt{\log(2/\delta)/(c(N-1))}$.
- $\dfrac{t}{\sigma_\eta^2} \geq \dfrac{\log(2/\delta)}{c(N-1)}$ gives $t \geq t_2 := \sigma_\eta^2\log(2/\delta)/(c(N-1))$.

Both conditions must hold simultaneously, so the threshold is the larger of the two: $t \geq \max(t_1, t_2)$. Using the elementary inequality $\max(t_1, t_2) \leq t_1 + t_2$ and writing $C := 1/c$:

$$|D^2 - \tau^2| \;\leq\; C\sigma_\eta^2\left(\sqrt{\frac{\log(2/\delta)}{N-1}} + \frac{\log(2/\delta)}{N-1}\right)$$

with probability at least $1-\delta$, which is Eq. 18. The two terms reflect the two tail regimes: the $\sqrt{\cdot}$ term dominates at small deviations (Gaussian-like behavior), the linear-in-log term dominates at large deviations (heavier tail behavior typical of quadratic forms of sub-Gaussians). □

The high-probability bound in Eq. 18 is honest about a fundamental limitation of small samples. To see the issue, specialise to the standard Gaussian case ($\sigma_\eta = \tau$) and plug in $\delta = 0.05$, $N = 4$:

$$|D^2 - \tau^2| \;\lesssim\; \tau^2\Big(\sqrt{\log(40)/3} + \log(40)/3\Big) \;\approx\; 2.3\,\tau^2.$$

The right-hand side is more than twice the very quantity we are trying to estimate. The bound therefore allows D^2 to lie anywhere in a wide neighbourhood of τ^2 on a single scenario. It cannot, on its own, certify that the persona panel is reliable.

Why this motivates the dual-pass gate. Theorem 3 formalises the gap that the dual-pass gate is designed to close. At $N{=}4$, D^2 is unbiased (Proposition 1(i)) but high-variance: it is the right object in expectation, but a single observation of it is too noisy to act on. We therefore need a *second* reliability signal that does not inherit D^2's small-sample limitation.

The dual-pass gate provides exactly this. Each IS pass averages over $K_{\text{half}} = 64$ candidate perturbations rather than $N{-}1 = 3$ persona deviations, so the inter-pass disagreement $V_r = (\delta^{(1)}_{\text{PT-IS}} - \delta^{(2)}_{\text{PT-IS}})^2$ concentrates around its target an order of magnitude faster than D^2 does. The two signals are also complementary rather than redundant: D^2 probes whether the persona panel agrees, while V_r probes whether the IS estimate stabilises across resamples. In the controller used for all results, D^2 enters only through the proposal scale σ; only V_r gates the correction. Theorems 3 and 2 together make this design rationale explicit: the gate exists because the within-panel variance estimator alone is provably insufficient at the panel size we use.

A3.3 Hölder stability with respect to backbone perturbations

We prove that two LLMs producing similar logits must produce similar corrections, regardless of how those logits arose.

The natural statement would be Lipschitz: $|\delta^\star_f - \delta^\star_{f'}| \le L\,\varepsilon$ for some constant L. We do not get this. The Kahneman-Tversky value function in Eq. 2 has derivative $\alpha z^{\alpha-1}$ on $z > 0$, which blows up as $z \to 0^+$ when $\alpha < 1$. A Lipschitz statement would require a global bound on the derivative; near the origin no such bound exists. The honest result is therefore α-Hölder continuity:

$$|\delta^\star_f - \delta^\star_{f'}| \;\le\; L\,\varepsilon^\alpha,$$

which is qualitatively the same statement (small input perturbations produce small output changes) with rate ε^α instead of ε. For $\alpha = 0.88$ and $\varepsilon \le 1$, $\varepsilon^{0.88}/\varepsilon \le 1.4$, so the quantitative loss is mild.

Theorem 4 (Hölder stability)**.** *Fix a scenario* $\mathbf{x}$ *and country* c*. For any LLM* f*, let* $\boldsymbol{\delta}_f = (\delta^f_{base}, \delta^f_1, \ldots, \delta^f_N)$ *denote the order-symmetrised base and persona logit gaps under* f*, and assume there is a uniform bound* $\|\boldsymbol{\delta}_f\|_\infty \le G$ *across the function class. For any two LLMs* f, f'*, define* $\varepsilon := \|\boldsymbol{\delta}_f - \boldsymbol{\delta}_{f'}\|_\infty$*. Assume further that the proposal scale* σ *is the same for* f *and* f' *and that the ESS guard is inactive in both passes (*$\delta^\star_m$ *is never set to zero). Then conditional on the IS perturbation draws* $\{\epsilon^{(m)}_k\}$*, shared by* f *and* f'*, with* $M := \max_{k,m} |\epsilon^{(m)}_k|$*, the DISCA corrections satisfy*

$$|\delta^\star_f - \delta^\star_{f'}| \;\le\; L(\sigma, \eta, s, \alpha, \kappa, M)\cdot \varepsilon^\alpha, \tag{20}$$

where

$$L(\sigma, \eta, s, \alpha, \kappa, M) \;=\; \frac{2\,M\,(1+\kappa)\,4^\alpha}{\eta\,\sigma^\alpha}\left(1 + \frac{8M^2}{s}\right). \tag{21}$$

Marginally over the IS draws, the same bound holds with M *replaced by* $\sigma\sqrt{2\log(4K_{half}/\delta_p)}$ *with probability at least* $1 - \delta_p$ *(Theorem 2).*

Proof. The proof tracks a perturbation through the seven stages of Algorithm 1, with each stage contributing an explicit Lipschitz or Hölder factor. We work in the ℓ_∞ norm throughout.

The IS perturbations $\{\epsilon^{(m)}_k\}$ are drawn from $\mathcal{N}(0, \sigma^2)$; with σ fixed by assumption, they do not depend on the LLM. In the controller σ depends on the persona spread and the ESS guard is a discontinuity, so the theorem describes the regime in which neither changes between f and f'. For the same scenario evaluated under f and f', we use the same IS draws (e.g., a fixed random seed). The proof therefore conditions on these draws, treating them as fixed inputs with $\max_{k,m} |\epsilon^{(m)}_k| =: M$. The marginal bound follows by combining the conditional bound with the high-probability bound on M from Theorem 2.

Stage A: Consensus. The map $\boldsymbol{\delta} \mapsto \bar\delta = \frac{1}{N}\sum_i \delta_i$ is 1-Lipschitz in ℓ_∞:

$$|\bar\delta_f - \bar\delta_{f'}| \;=\; \left|\tfrac{1}{N}\sum_{i=1}^N (\delta^f_i - \delta^{f'}_i)\right| \;\le\; \tfrac{1}{N}\sum_{i=1}^N |\delta^f_i - \delta^{f'}_i| \;\le\; \varepsilon.$$

At the entry to Stage B, we therefore have $|\delta^f_{\text{base}} - \delta^{f'}_{\text{base}}|, |\delta^f_i - \delta^{f'}_i|, |\bar\delta^f - \bar\delta^{f'}| \le \varepsilon$.

Stage B: Cohort and consensus gains. The gain $g_{i,k}$ in Eq. 1 is $g_{i,k} = |\delta_{\text{base}} - \delta_i| - |\tilde{\delta}_k - \delta_i|$, with $\tilde{\delta}_k = \bar{\delta} + \epsilon_k$. We bound $|g^f_{i,k} - g^{f'}_{i,k}|$ by triangle inequality on the difference of the two absolute-value terms, then by reverse-triangle inequality $\big||a| - |b|\big| \le |a - b|$ on each:

$$\begin{aligned}|g^f_{i,k} - g^{f'}_{i,k}| &\le \big||\delta^f_{\text{base}} - \delta^f_i| - |\delta^{f'}_{\text{base}} - \delta^{f'}_i|\big| + \big||\tilde{\delta}^f_k - \delta^f_i| - |\tilde{\delta}^{f'}_k - \delta^{f'}_i|\big| \\ &\le |(\delta^f_{\text{base}} - \delta^f_i) - (\delta^{f'}_{\text{base}} - \delta^{f'}_i)| + |(\bar{\delta}^f - \delta^f_i) - (\bar{\delta}^{f'} - \delta^{f'}_i)|.\end{aligned}$$

The IS perturbation ϵ_k cancels in the second term since it is LLM-independent. Each remaining term is at most 2ε by triangle inequality applied to the difference of two coordinates of $\boldsymbol{\delta}$. Therefore

$$|g^f_{i,k} - g^{f'}_{i,k}| \;\le\; 4\varepsilon, \qquad |g^f_{\text{cons},k} - g^{f'}_{\text{cons},k}| \;\le\; 4\varepsilon.$$

Stage C: Prospect-Theory value function. We claim the Kahneman-Tversky value function in Eq. 2 is α-Hölder with constant $1 + \kappa$:

$$|v(z) - v(z')| \;\le\; (1 + \kappa)\,|z - z'|^\alpha \qquad \forall z, z' \in \mathbb{R}. \tag{22}$$

Lemma 1 (subadditivity)**.** *For $\alpha \in (0, 1]$ and $a, b \ge 0$: $(a + b)^\alpha \le a^\alpha + b^\alpha$.*

Proof. The boundary cases ($\alpha = 1$, $a = 0$, or $b = 0$) give equality. For $\alpha \in (0, 1)$ and $a, b > 0$, divide both sides by $(a+b)^\alpha$ and set $u := a/(a+b)$, $v := b/(a+b)$, so $u, v \in (0, 1)$ and $u + v = 1$. The inequality reduces to $u^\alpha + v^\alpha \ge 1$.

For any $x \in (0, 1)$ and $\alpha \in (0, 1)$: $x^\alpha = x \cdot x^{\alpha-1}$, and $x^{\alpha-1} > 1$ (since $x < 1$ and $\alpha - 1 < 0$), so $x^\alpha > x$. Therefore $u^\alpha + v^\alpha > u + v = 1$. □

Corollary 5. *For $\alpha \in (0, 1]$ and any reals a, b: $\big||a|^\alpha - |b|^\alpha\big| \le |a - b|^\alpha$. WLOG $|a| \ge |b|$; the lemma applied to $|a| = (|a| - |b|) + |b|$ gives $|a|^\alpha \le (|a| - |b|)^\alpha + |b|^\alpha$, so $|a|^\alpha - |b|^\alpha \le (|a| - |b|)^\alpha \le |a - b|^\alpha$ by reverse triangle inequality and monotonicity of x^α.*

We now verify Eq. 22 by cases on the signs of z, z':

- *Both $z, z' \ge 0$:* $|v(z) - v(z')| = |z^\alpha - z'^\alpha| \le |z - z'|^\alpha$ by the corollary, the constant is 1.
- *Both $z, z' < 0$:* $|v(z) - v(z')| = \kappa\big||-z|^\alpha - |-z'|^\alpha\big| \le \kappa\,|z - z'|^\alpha$, the constant is κ.
- *Cross-sign, $z \ge 0$ and $z' < 0$:* writing $a := z \ge 0$ and $b := -z' > 0$, we have $|z - z'| = a + b$ and $|v(z) - v(z')| = a^\alpha + \kappa b^\alpha$. Using $a^\alpha, b^\alpha \le (a + b)^\alpha$ (monotonicity of x^α):

$$|v(z) - v(z')| = a^\alpha + \kappa b^\alpha \le (1 + \kappa)(a + b)^\alpha = (1 + \kappa)|z - z'|^\alpha.$$

The maximum of the three case-constants is $1 + \kappa$ (since $\kappa \ge 1$), proving Eq. 22.

Rescaling input by σ and combining with Stage B:

$$\big|v(g^f_{i,k}/\sigma) - v(g^{f'}_{i,k}/\sigma)\big| \;\le\; \frac{1 + \kappa}{\sigma^\alpha}\,|g^f_{i,k} - g^{f'}_{i,k}|^\alpha \;\le\; \frac{(1 + \kappa)\,4^\alpha}{\sigma^\alpha}\,\varepsilon^\alpha \;=:\; C_v\,\varepsilon^\alpha.$$

Stage D: Total utility. $U_{\text{total}}(\epsilon_k)$ in Eq. 3 is a convex combination: weights $(1 - \lambda_{\text{coop}})/N$ on each of the N cohort terms and λ_{coop} on the consensus term, summing to $(1 - \lambda_{\text{coop}}) + \lambda_{\text{coop}} = 1$. Convex combinations preserve ℓ_∞-bounds: if $|y^f_j - y^{f'}_j| \le B$ for each summand and $w_j \ge 0$ with $\sum_j w_j = 1$, then $|\sum_j w_j y^f_j - \sum_j w_j y^{f'}_j| \le \sum_j w_j |y^f_j - y^{f'}_j| \le B$. Applied to the $N + 1$ component v-values bounded by $C_v \varepsilon^\alpha$ from Stage C:

$$|U^f_k - U^{f'}_k| \;\le\; C_v\,\varepsilon^\alpha \qquad \forall k.$$

Stage E: Importance-sampling aggregator. The map $\mathbf{U} \mapsto \delta_{\text{IS}} = \sum_k w_k \epsilon_k$ uses softmax weights $w_k = \exp(U_k/\eta)/\sum_l \exp(U_l/\eta)$. The standard softmax Jacobian, by quotient rule, is

$$\frac{\partial w_j}{\partial U_k} \;=\; \frac{1}{\eta}\,w_j\,(\mathbb{I}[j = k] - w_k).$$

Since the IS draws $\{\epsilon_j\}$ are constants (LLM-independent), the chain rule gives

$$\frac{\partial \delta_{\text{IS}}}{\partial U_k} \;=\; \sum_j \epsilon_j \frac{\partial w_j}{\partial U_k} \;=\; \frac{1}{\eta}\left(w_k \epsilon_k - w_k \sum_j w_j \epsilon_j \right) \;=\; \frac{w_k}{\eta}(\epsilon_k - \delta_{\text{IS}}),$$

using $\sum_j w_j \epsilon_j = \delta_{\text{IS}}$ in the last step. Both $|\epsilon_k|$ and $|\delta_{\text{IS}}|$ are bounded by M (the latter as a convex combination of $\{\epsilon_j\}$), so $|\epsilon_k - \delta_{\text{IS}}| \le 2M$. Summing the absolute partials:

$$\sum_k \left| \frac{\partial \delta_{\text{IS}}}{\partial U_k} \right| \;\le\; \frac{2M}{\eta} \sum_k w_k \;=\; \frac{2M}{\eta}.$$

By the multivariable mean value theorem and ℓ_1/ℓ_∞ duality, $\sum_k |\partial f/\partial U_k| \le L$ implies f is L-Lipschitz in ℓ_∞. Therefore δ_{IS} is $(2M/\eta)$-Lipschitz in $\mathbf{U}$, and combining with Stage D:

$$|\delta_{\text{IS}}^{f} - \delta_{\text{IS}}^{f'}| \;\le\; \frac{2M}{\eta} \, \|\mathbf{U}^f - \mathbf{U}^{f'}\|_\infty \;\le\; \frac{2MC_v}{\eta} \, \varepsilon^\alpha.$$

Stage F: Reliability gate. Let $D := \delta_{\text{IS}}^{(1)} - \delta_{\text{IS}}^{(2)}$, so $V_r = D^2$. Using the factorisation $|a^2 - b^2| = |a+b|\,|a-b|$:

$$|V_r^f - V_r^{f'}| \;=\; |D^f + D^{f'}|\,|D^f - D^{f'}|.$$

Each $|D|$ is bounded by $|\delta_{\text{IS}}^{(1)}| + |\delta_{\text{IS}}^{(2)}| \le 2M$, so $|D^f + D^{f'}| \le 4M$. The difference $|D^f - D^{f'}|$ involves both passes' Stage-E outputs:

$$|D^f - D^{f'}| \;\le\; \left|\delta_{\text{IS}}^{(1),f} - \delta_{\text{IS}}^{(1),f'}\right| + \left|\delta_{\text{IS}}^{(2),f} - \delta_{\text{IS}}^{(2),f'}\right| \;\le\; \frac{4MC_v}{\eta}\varepsilon^\alpha,$$

where each pass contributes a Stage E bound of $(2MC_v/\eta)\varepsilon^\alpha$ and they add. Combining:

$$|V_r^f - V_r^{f'}| \;\le\; 4M \cdot \frac{4MC_v}{\eta}\varepsilon^\alpha \;=\; \frac{16M^2C_v}{\eta} \, \varepsilon^\alpha.$$

The gate $r = \exp(-V_r/s)$ from Eq. 5 is $1/s$-Lipschitz in V_r on $[0,\infty)$, since $|d/dx\, e^{-x/s}| = (1/s)e^{-x/s} \le 1/s$ for $x \ge 0$. Therefore

$$|r^f - r^{f'}| \;\le\; \frac{1}{s} \cdot \frac{16M^2C_v}{\eta} \, \varepsilon^\alpha \;=\; \frac{16M^2C_v}{s\eta} \, \varepsilon^\alpha.$$

Stage G: Final correction. Define $A := (\delta_{\text{IS}}^{(1)} + \delta_{\text{IS}}^{(2)})/2$. Each per-pass output is bounded by M in absolute value (Stage E), so $|A| \le M$ by triangle inequality. The difference between LLMs:

$$|A^f - A^{f'}| \;\le\; \frac{1}{2}\Big(\left|\delta_{\text{IS}}^{(1),f} - \delta_{\text{IS}}^{(1),f'}\right| + \left|\delta_{\text{IS}}^{(2),f} - \delta_{\text{IS}}^{(2),f'}\right|\Big) \;\le\; \frac{2MC_v}{\eta} \, \varepsilon^\alpha,$$

where the $1/2$ in A cancels the factor of 2 from summing both passes.

Apply the product rule to $\delta^\star = r \cdot A$, adding and subtracting the cross term $r^f A^{f'}$:

$$|\delta_f^\star - \delta_{f'}^\star| \;=\; |r^f A^f - r^{f'} A^{f'}| \;\le\; |r^f| \cdot |A^f - A^{f'}| + |A^{f'}| \cdot |r^f - r^{f'}|.$$

Substituting $|r^f| \le 1$ (from Eq. 5, $r \in (0,1]$), $|A^{f'}| \le M$, the bound on $|A^f - A^{f'}|$, and the Stage F bound on $|r^f - r^{f'}|$:

$$|\delta_f^\star - \delta_{f'}^\star| \le 1 \cdot \frac{2MC_v}{\eta}\varepsilon^\alpha + M \cdot \frac{16M^2C_v}{s\eta}\varepsilon^\alpha \;=\; \frac{2MC_v}{\eta}\left(1 + \frac{8M^2}{s}\right)\varepsilon^\alpha.$$

Substituting $C_v = (1+\kappa)\,4^\alpha/\sigma^\alpha$ yields Eq. 21. The marginal statement follows from the high-probability bound on M in Theorem 2. □

This result shows that DISCA correction is stable in the LLM, in the structural sense: it is a continuous function of persona logit gaps, with the explicit (if loose) modulus of Eq. 21. Two LLMs producing similar persona logits must produce similar corrections and the modulus depends only on hyperparameters $(\sigma, \eta, s, \alpha, \kappa)$, not on the LLM. The bound concerns $\delta^\star$ only; the anchor and the country prior are not covered.

The result has limitation to read carefully. It is conditional on the boundedness assumption $\|\boldsymbol{\delta}_f\|_\infty \leq G$. Models with collapsed logit entropy (App. A7) fall outside this regime: the empirical observation that those models degrade under DISCA is consistent with ε being effectively unbounded for them, not a contradiction of the theorem. This is a complementary role to Theorem 2: that theorem gives a numerical bound on $|\delta^\star|$ from the proposal scale and budget $(\sigma, K_{\text{half}})$; this theorem gives a loose-but-structural continuity bound across LLMs, holding under the boundedness regime. Together they characterise both axes of DISCA's behaviour.

A4 Full Per-Country Results

Table 7 lists, for every country and backbone, the vanilla and DISCA misalignment scores, the relative change, the Jensen-Shannon distance and Pearson r; the macro means match Table 2.

A5 Per-Dimension Error Analysis



Figure 2: **Per-dimension DISCA improvement across the seven headline backbones.** Each cell is the macro-averaged (over 20 countries) reduction in per-dimension error (percentage points): $\Delta = |\text{vanilla} - \text{human}| - |\text{DISCA} - \text{human}|$. Positive (blue) means DISCA helped on that dimension; negative (red) means it hurt. Averaged over the seven backbones, DISCA lowers the error most on **Utilitarianism** (8.4 pp), followed by Species (3.8), Gender (3.2) and Social Value (3.1); it raises the Age error (2.2 pp). The companion worst-error analysis (Table 8) reports the dominant bottleneck dimension of each backbone under DISCA. Cells with $\Delta > 5$ pp shown in bold.

Table 8: Per-dimension worst-error analysis under DISCA across the 20-country panel for the seven headline backbones. For each model we report the dimension that is most often the single worst, its frequency, and its mean absolute error over the countries where it is worst. The rightmost columns give the *second*-most-frequent worst dimension, to show whether errors are concentrated or spread.

Model	Worst dim	Freq.	Mean err (pp)	2nd worst dim	Freq.
Llama-3.3-70B	Utilitarianism	14/20	49.6	Social Value	6/20
Phi-3.5-mini	Utilitarianism	18/20	41.1	Species	2/20
Qwen3-VL-8B	Social Value	18/20	34.3	Utilitarianism	2/20
Qwen2.5-7B	Social Value	14/20	24.6	Species	3/20
Magistral-24B	Social Value	14/20	22.4	Age	3/20
Phi-4	Social Value	9/20	22.2	Age	7/20
Gemma-4-E2B	Age	10/20	27.6	Utilitarianism / Species	4/20 each

Table 7: **Per-country DISCA results (20 countries).** Countries are grouped by geographic region. Vanilla and DISCA misalignment scores (MIS; ↓), relative MIS improvement (%), and DISCA Jensen-Shannon distance (JSD; ↓) / Pearson r (↑). Seven headline models (nominal parameter count, descending).

		Llama-3.3-70B					**Magistral-Small (24B)**					**Phi-4 (14B)**					**Qwen3-VL-8B**					**Qwen2.5-7B**					**Phi-3.5-mini (3.8B)**					**Gemma-4-E2B (2B)**				
ISO	**Region**	MIS_{van}	MIS_{DISCA}	$\%\Delta_{MIS}$	JSD_D	r_D	MIS_{van}	MIS_{DISCA}	$\%\Delta_{MIS}$	JSD_D	r_D	MIS_{van}	MIS_{DISCA}	$\%\Delta_{MIS}$	JSD_D	r_D	MIS_{van}	MIS_{DISCA}	$\%\Delta_{MIS}$	JSD_D	r_D	MIS_{van}	MIS_{DISCA}	$\%\Delta_{MIS}$	JSD_D	r_D	MIS_{van}	MIS_{DISCA}	$\%\Delta_{MIS}$	JSD_D	r_D	MIS_{van}	MIS_{DISCA}	$\%\Delta_{MIS}$	JSD_D	r_D
ARG	Americas	1.015	.891	+12.2	.178	−0.456	.364	.342	+6.0	.046	.601	.463	.389	+16.1	.046	.412	.520	.338	+35.0	.083	.508	.469	.377	+19.7	.064	.196	.770	.645	+16.1	.159	−0.498	.423	.434	-2.4	.050	.213
BRA	Americas	.521	.458	+12.1	.119	.222	.412	.319	+22.5	.051	.333	.274	.299	-9.3	.061	.471	.529	.375	+29.1	.101	.212	.579	.411	+29.1	.081	−0.205	.531	.649	-22.1	.173	−0.743	.416	.404	+2.9	.074	.211
COL	Americas	1.041	.931	+10.5	.189	−0.594	.413	.369	+10.7	.055	.386	.487	.438	+10.1	.057	.093	.559	.390	+30.1	.093	.161	.508	.417	+17.8	.070	−0.096	.788	.678	+14.0	.165	−0.602	.464	.460	+0.9	.051	−0.011
MEX	Americas	1.037	.909	+12.3	.179	−0.385	.349	.333	+4.6	.040	.707	.485	.420	+13.4	.044	.512	.517	.350	+32.3	.081	.529	.467	.388	+17.0	.062	.252	.782	.662	+15.4	.159	−0.494	.430	.456	-5.9	.041	.503
USA	Americas	.873	.578	+33.8	.121	.266	.448	.374	+16.5	.042	.731	.497	.243	+51.1	.058	.723	.677	.487	+28.0	.110	.422	.483	.383	+20.8	.075	.709	.645	.579	+10.2	.110	−0.380	.545	.517	+5.3	.060	.618
DEU	Europe	.715	.643	+10.1	.189	.101	.264	.345	-30.3	.045	.724	.302	.255	+15.6	.060	.779	.318	.306	+3.7	.081	.568	.308	.415	-34.6	.069	.144	.569	.555	+2.5	.146	−0.403	.506	.416	+17.8	.053	.617
GBR	Europe	.873	.598	+31.4	.141	.125	.442	.384	+13.2	.048	.674	.503	.319	+36.6	.082	.594	.677	.524	+22.6	.128	.230	.483	.399	+17.4	.082	.620	.654	.588	+10.1	.114	−0.392	.556	.502	+9.8	.062	.342
ROU	Europe	.851	.593	+30.3	.147	.197	.446	.363	+18.7	.043	.708	.504	.385	+23.5	.058	.570	.659	.535	+18.8	.123	.162	.442	.330	+25.2	.051	.785	.671	.577	+14.0	.117	−0.352	.547	.482	+11.8	.045	.644
SRB	Europe	.862	.605	+29.9	.150	.157	.463	.372	+19.8	.045	.670	.500	.385	+22.9	.057	.564	.676	.567	+16.2	.127	.075	.453	.303	+33.1	.051	.773	.677	.601	+11.1	.124	−0.463	.548	.502	+8.4	.047	.573
CHN	E. Asia	.901	.748	+16.9	.183	−0.025	.367	.293	+20.2	.050	.711	.407	.214	+47.4	.054	.781	.569	.416	+27.0	.121	.269	.437	.356	+18.6	.084	.405	.738	.559	+24.3	.162	−0.041	.500	.393	+21.5	.039	.767
JPN	E. Asia	.521	.471	+9.5	.107	.249	.341	.265	+22.3	.054	.709	.395	.195	+50.6	.051	.671	.408	.316	+22.4	.090	.453	.296	.356	-20.1	.047	.535	.454	.455	-0.3	.063	−0.233	.442	.460	-4.0	.062	−0.357
IDN	SE. Asia	.688	.557	+19.0	.092	.071	.283	.344	-21.5	.059	.511	.459	.274	+40.3	.059	.595	.408	.402	+1.5	.091	.318	.480	.402	+16.1	.058	.251	.492	.484	+1.6	.120	−0.769	.417	.443	-6.2	.044	.572
MMR	SE. Asia	.872	.652	+25.2	.185	−0.031	.496	.365	+26.5	.062	.457	.482	.357	+26.0	.057	.570	.695	.563	+18.9	.139	−0.066	.461	.302	+34.3	.062	.590	.685	.590	+13.8	.135	−0.623	.488	.481	+1.4	.061	.372
MYS	SE. Asia	.840	.589	+29.9	.155	.003	.443	.313	+29.4	.043	.683	.459	.328	+28.5	.053	.614	.656	.504	+23.1	.128	.097	.445	.274	+38.5	.050	.763	.631	.535	+15.1	.118	−0.426	.484	.430	+11.2	.039	.676
THA	SE. Asia	.814	.577	+29.2	.152	.125	.406	.295	+27.3	.041	.738	.456	.313	+31.4	.047	.685	.619	.480	+22.5	.112	.231	.414	.221	+46.6	.046	.841	.619	.520	+16.1	.112	−0.283	.470	.436	+7.2	.040	.750
VNM	SE. Asia	.845	.740	+12.4	.142	−0.045	.324	.347	-7.2	.050	.589	.407	.336	+17.3	.064	.596	.444	.420	+5.5	.087	.339	.501	.468	+6.6	.060	−0.018	.481	.507	-5.4	.067	−0.143	.468	.544	-16.1	.048	.566
BGD	S. Asia	.864	.628	+27.3	.162	.029	.450	.361	+19.6	.047	.625	.489	.368	+24.7	.048	.607	.674	.587	+12.9	.133	−0.116	.461	.284	+38.3	.055	.695	.666	.565	+15.3	.115	−0.426	.519	.496	+4.4	.052	.435
KGZ	C. Asia	.849	.585	+31.1	.149	.176	.460	.353	+23.2	.048	.633	.499	.393	+21.1	.058	.476	.661	.574	+13.1	.133	−0.017	.443	.249	+43.7	.039	.829	.660	.572	+13.3	.119	−0.405	.518	.473	+8.7	.049	.600
IRN	W. Asia	1.039	.852	+18.0	.159	−0.294	.427	.401	+6.1	.055	.637	.397	.530	-33.5	.087	.055	.507	.506	+0.2	.090	.057	.373	.466	-25.1	.067	.252	.511	.445	+12.9	.056	.766	.514	.630	-22.6	.088	.408
ETH	Africa	.967	.747	+22.7	.172	−0.077	.462	.472	-2.0	.054	.655	.615	.485	+21.2	.041	.808	.723	.684	+5.4	.141	−0.095	.558	.438	+21.6	.059	.676	.734	.661	+10.0	.114	−0.027	.645	.612	+5.0	.057	.628
Mean		**.849**	**.668**	**+21.4**	**.154**	**−0.009**	**.403**	**.350**	**+13.1**	**.049**	**.624**	**.454**	**.346**	**+23.7**	**.057**	**.559**	**.575**	**.466**	**+18.9**	**.110**	**.217**	**.453**	**.362**	**+20.1**	**.062**	**.450**	**.638**	**.571**	**+10.4**	**.122**	**−0.347**	**.495**	**.478**	**+3.4**	**.053**	**.456**

Legend. MIS_{van}/MIS_{DISCA}: vanilla vs. DISCA ℓ_2 misalignment; $\%\Delta_{MIS}$: relative improvement in MIS; JSD_D and r_D: Jensen-Shannon distance and Pearson correlation under DISCA. **Region groups:** Americas (ARG, BRA, COL, MEX, USA); Europe (DEU, GBR, ROU, SRB); E. Asia (CHN, JPN); SE. Asia (IDN, MMR, MYS, THA, VNM); S. Asia (BGD); C. Asia (KGZ); W. Asia (IRN); Africa (ETH).

Table 9: Phi-4 per-country MIS (↓): the best-performing model in absolute terms. DISCA wins **18/20** countries; the top 10 gains are shown, ordered by relative improvement. The two losses (BRA, IRN) start from a vanilla MIS below the macro mean.

Country	Vanilla	DISCA	Gain (%)	Country	Vanilla	DISCA	Gain (%)
USA	0.497	**0.243**	+51.1	JPN	0.395	**0.195**	+50.6
CHN	0.407	**0.214**	+47.4	IDN	0.459	**0.274**	+40.3
GBR	0.503	**0.319**	+36.6	THA	0.456	**0.313**	+31.4
MYS	0.459	**0.328**	+28.5	MMR	0.482	**0.357**	+26.0
BGD	0.489	**0.368**	+24.7	ROU	0.504	**0.385**	+23.5
Losses: BRA $0.274 \rightarrow 0.299$ (-9.3%)				IRN $0.397 \rightarrow 0.530$ (-33.5%)			
Macro mean: $0.454 \rightarrow$ **0.346**				+23.7%, 18/20 wins, mean $r = +0.56$			

Three structural insights emerge from these tables.

First, **the bottleneck dimension shifts with model quality**. Weaker models (Llama-3.3-70B, Phi-3.5-mini) are dominated by **Utilitarianism**. Even after DISCA they predict about 30 to 35% utilitarian preference where humans range from 68 to 80%, a mean error of about 41 to 50 pp that no persona-based correction closes. Better-calibrated models (Phi-4, Magistral, Qwen2.5-7B) have already resolved Utilitarianism and Species, shifting the bottleneck to **SocialValue** and **Age**, with mean errors of about 22 to 25 pp under DISCA. On these three models DISCA lowers the Social Value error but raises the Age error.

Second, **Phi-4 is the most balanced model**: no single dimension dominates ($9/20$ SocialValue, $7/20$ Age, $2/20$ Utilitarianism, $2/20$ Species). This balance, not raw scale, explains why it achieves the lowest absolute MIS despite being $5\times$ smaller than Llama-3.3-70B.

Third, the per-country Phi-4 table (Table 9) confirms that **gains are large and geographically broad**: 51% on USA and Japan, 47% on China, 40% on Indonesia. The two losses (Brazil, Iran) are both in countries whose vanilla MIS was already low.

A5.1 Diagnosing negative Pearson r

Pearson r is negative in 36 of the 140 backbone-country cells under DISCA, against 41 under vanilla decoding (Table 10). The negative cells concentrate in the two backbones whose errors are dominated by Utilitarianism: Phi-3.5-mini (19 cells) and Llama-3.3-70B (8 cells). Phi-4 and Magistral-24B have none. In 24 of the 36 cells r was already negative under vanilla decoding, and DISCA lowers MIS in 32 of the 36 cells. The two metrics answer different questions. MIS measures the ℓ_2 distance on the six-dimensional AMCE vector, while r measures whether the model ranks the six attributes in the human order. A correction can shrink the distance while one large error still sets the ranking. Utilitarianism is that error here: in every one of the 36 cells the model under-predicts the human preference for saving more lives, and setting only the Utilitarianism coordinate to its human value makes r positive in 31 of them.

Table 10: Number of the 20 countries with negative Pearson r between model and human AMCE vectors, per backbone (seed-42 headline runs).

Backbone	Vanilla	DISCA
Llama-3.3-70B	7	8
Magistral-24B	0	0
Phi-4 (14B)	10	0
Qwen3-VL-8B	4	4
Qwen2.5-7B	2	3
Phi-3.5-mini (3.8B)	15	19
Gemma-4-E2B (2B)	3	2
Total (of 140)	41	36

A6 Geometric and Per-Country Visualizations

Figure 3 shows the AMCE vectors of Llama-3.3-70B in a two-component PCA plane, the geometric view of the gain reported in §4.1. Figure 4 shows the gain of DISCA over vanilla decoding for each of the 20 countries, averaged across the seven headline backbones.

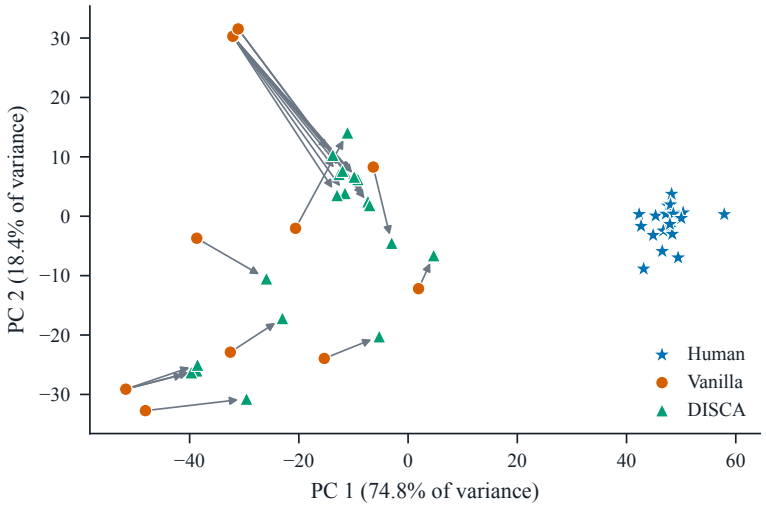


Figure 3: **DISCA pulls model preferences toward the human ones.** Two-component PCA of the six-dimensional human, vanilla and DISCA AMCE vectors of Llama-3.3-70B for all 20 countries (93.2% of the variance). Arrows trace each country from vanilla to DISCA. The mean distance to the human point falls from 83.3 to 64.6 in this plane and from 84.9 to 66.8 in all six dimensions, on the percentage-point scale (App. A15).

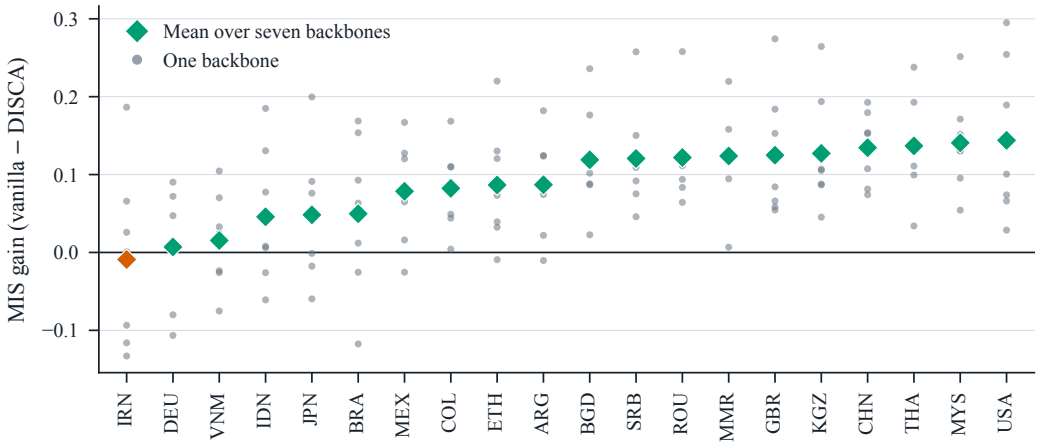


Figure 4: **DISCA gain per country.** MIS gain (vanilla minus DISCA, higher is better) for each of the 20 countries, sorted by the mean over the seven headline backbones (diamonds); grey dots are the individual backbones. **19 of 20 countries see a positive mean gain**; only Iran is slightly negative, and Germany and Vietnam are close to zero. The largest mean gains are in the USA, Malaysia, Thailand and China, so the gain is not confined to Western countries.

A7 Broader Model Landscape

The seven models in Table 2 were selected from a broader sweep of **22 open-weight models**: the 13 with a positive macro gain are listed below, and the 9 with no gain are omitted. Table 11 summarises the full landscape, ordered by macro improvement versus vanilla.

Table 11: Broader model sweep (5-country prototyping panel: BRA, CHN, DEU, JPN, USA). Models above the line show positive macro gains; those below show degradation. Bolded rows appear in the main paper's 20-country evaluation.

Model	Params	Mean MIS ↓	vs. van. (%)	Win/5
Phi-4	14B	**0.246**	+34.4	4/5
Llama-3.3-70B (4-bit)	70B	0.524	+25.9	5/5
Qwen3-VL-8B	8B	0.381	+23.7	4/5
Llama-3.1-8B (4-bit)	8B	0.455	+16.9	5/5
Magistral-Small-2509	24B	0.317	+13.6	4/5
Gemma-4-E2B	2B	0.426	+11.6	4/5
Qwen2.5-7B (bf16)	7B	0.385	+8.2	3/5
Phi-3.5-mini	3.8B	0.545	+7.3	4/5
Mistral-7B-v0.3	7B	0.428	+5.9	5/5
Qwen3.5-0.8B	0.8B	0.471	+3.2	4/5
Gemma-7B	7B	0.432	+2.5	3/5
Gemma-3-270M	270M	0.462	+1.3	4/5
Llama-3.2-1B	1B	0.476	+1.1	1/5
9 additional models showed ≤ 0 *macro gain (omitted)*				

Several patterns inform the main paper's model selection. *Success correlates with logit well-behavedness, not raw scale:* Phi-4 (14B) leads while Gemma-4-31B (31B) and Qwen3-Coder-30B (30B) degrade, which suggests that DISCA needs a well-conditioned decision surface, not just more parameters. *Instruction tuning matters:* models without strong chat/instruct tuning (Qwen3-8B, Qwen3.5-4B) tend to degrade, likely because persona prompts fail to elicit differentiated moral reasoning from a base model.

A8 Factual Cultural QA Evaluation (BLEnD)

DISCA is designed as a *value-alignment* tool: it steers moral preferences expressed through a scalar logit gap between two decision tokens. A natural question is whether the same persona-disagreement signal transfers to *factual cultural knowledge*, where the correct answer is a specific token or span in a softmax over a vocabulary of about 100k entries (Phi-4). This is a very different objective. We evaluate on the BLEnD benchmark [Myung et al., 2024] (52.6k question-answer pairs, 16 countries and regions, 13 languages) to probe this boundary. Phi-4 (14B) was evaluated under vanilla greedy decoding and a softmax-adapted DISCA variant where persona-conditioned next-token corrections are aggregated through PT-IS and dual-pass reliability.

Where DISCA transfers. **Seven of sixteen** countries see a positive SEM-B gain under the softmax-adapted DISCA variant (Table 12). The wins fall into two regimes. The largest gains land on *high-baseline* countries: the United Kingdom ($+7.45$ pp), the United States ($+4.55$), Mexico, Spain, and Indonesia. There the vanilla baseline already generates fluent, culturally appropriate text, and the persona signal refines an already coherent decision surface. A second pocket of smaller positive transfer appears at the *low-baseline* end, in West Java ($+2.40$ pp on a vanilla baseline of 26.64%) and Azerbaijan ($+0.43$ on 28.78%). The persona signal can therefore also help some lower-resource panels even when factual recall is weak.

Table 12: BLEnD per-country gains (Phi-4 14B). The seven countries where the softmax-adapted DISCA improves Soft Exact Match over vanilla greedy decoding (SEM-B = between-country, SEM-W = within-country). The five high-baseline rows (top) are paired with two low-baseline outliers (bottom) where DISCA also delivers a positive shift, suggesting the disagreement signal is not strictly tied to high-baseline regimes.

Country	Vanilla (%)		Δ DISCA (pp)	
	SEM-B	SEM-W	SEM-B	SEM-W
UK	75.66	66.69	+7.45	+8.78
US	82.43	74.32	+4.55	+3.80
Mexico	72.65	61.75	+2.09	+2.76
Spain	72.71	63.12	+1.92	+2.89
Indonesia	73.75	60.92	+1.67	+3.39
West Java	26.64	20.73	+2.40	+3.81
Azerbaijan	28.78	24.92	+0.43	+0.48

The value-fact boundary. Aggregate SEM-B across all 16 countries drops by 4.4 pp because the remaining nine countries regress under DISCA. They are mostly mid- and low-baseline cases such as Greece, Iran, and Assam. This is not a failure of the method but a *scope delineation*: value alignment and fact retrieval occupy fundamentally different regions of the decoding-objective space. In value alignment, the correct answer is a cultural preference encoded in a scalar logit gap between two options, which is exactly the space DISCA's IS stage searches. In factual QA, the correct answer is a single token in a vocabulary of about 100k entries; injecting perturbations into this high-dimensional softmax adds noise that drowns the cultural signal when the baseline is mid-range. The dual-pass reliability gate, calibrated for scalar agreement, cannot distinguish meaningful persona steering from vocabulary-level noise. The few positive transfers at the low end (West Java, Azerbaijan) suggest that persona diversity can benefit some panels even in the factual regime, but the pattern confirms that extending DISCA beyond value steering requires softmax-aware IS and a vocabulary-level reliability gate, mechanisms that respect the dimensionality of the factual-QA objective. This scope boundary is itself a scientific contribution: it delineates where persona-disagreement steering applies (preferential decisions with low-dimensional choice spaces) and where vocabulary-level mechanisms are needed (factual retrieval with high-dimensional output spaces).

A9 Post-Hoc Diagnostics

This appendix collects two diagnostics computed post-hoc from the per-scenario outputs of the DISCA runs (no model reload required): the seeds behind each result (§A9.1) and the scenario-level link between within-panel disagreement and the applied correction (§A9.2). Shape diagnostics complementing the ℓ_2-based MIS appear in App. A5.1.

A9.1 Seeds

Every headline number in Table 2 and every per-country table comes from one run per backbone with seed 42. The two Qwen2.5 backbones (7B and 14B) were also run on the full 20-country panel with seeds 42, 101 and 2026. These runs supply the same-logit controls (Table 4), the GlobalOpinionQA DISCA average (Table 18; there the three seeds rerun the controller on one set of model outputs), and the persona-tail statistic of §5. That statistic is the share of scenarios in which the corrected gap ends further from at least one persona's own gap than the vanilla gap was: 51.4% on Qwen2.5-7B and 53.3% on Qwen2.5-14B, averaged over the three seeds.

A9.2 Scenario-level conditioning of the DISCA correction

Proposition 1 formalises DISCA as a variance-aware shrinkage estimator whose strength is governed by within-panel disagreement $D^2 = \frac{1}{N-1}\sum_i(\delta_i - \bar{\delta})^2$: substituting the unbiased estimate $\widehat{\tau^2} = D^2$ into the oracle factor $\gamma^\star = \Delta_h^2/(\Delta_h^2 + \tau^2/N)$ yields a shrinkage weight that is monotone-decreasing

in D^2, so D^2 controls the *ratio* $|\delta^\star|/|\Delta|$ rather than $|\delta^\star|$ in isolation. A simpler but operationally important question, complementary to the proposition, is whether DISCA's applied correction is in fact driven by disagreement *at the scenario level*, or whether it collapses into a per-country offset. Using Qwen2.5-7B as the backbone, we log *per-scenario* $(D^2, |\delta^\star|)$ pairs from the full DISCA pipeline across 5 countries (USA, JPN, DEU, VNM, ETH) and 310 MultiTP scenarios per country ($n = 1{,}550$ total), and compute Pearson's correlation between $\log_{10} D^2$ and $|\delta^\star|$.

Result. The two quantities are positively and significantly correlated, $r = +0.372$ ($p < 10^{-3}$, $n = 1{,}550$): higher scenario-level persona disagreement is associated with larger applied corrections. This is consistent with the scenario-conditioned shrinkage policy formalised by Proposition 1, with the caveat that the proposition governs the ratio $|\delta^\star|/|\Delta|$ rather than the raw magnitude. A positive raw correlation arises when $|\Delta|$ also co-varies with D^2 across scenarios, which is the typical case when both are driven by the same underlying spread of persona logit gaps. At the country level (Table 13), the extremes match expectation: ETH, with the lowest mean variance (0.009), receives the smallest mean correction (0.009); USA, with the highest mean variance (0.086), receives the largest (0.015). The three middle countries (JPN, DEU, VNM) sit in a narrow band (0.012 to 0.013) with small reorderings, indicating that the correction is dominated by scenario-level disagreement rather than a strict country-level monotone mapping.

Table 13: Per-country mean inter-persona variance $\overline{D^2}$ and mean correction magnitude $\overline{|\delta^\star|}$ ($n = 310$ each). Countries are ordered by mean variance. The two extremes (ETH, USA) match expectation; the three middle countries sit in a narrow band, showing the correction is set primarily at the scenario level rather than by a strict country-level monotone mapping.

Country	n	$\overline{D^2}$	$\overline{\lvert\delta^\star\rvert}$
ETH	310	0.009	0.009
JPN	310	0.052	0.013
DEU	310	0.055	0.012
VNM	310	0.070	0.013
USA	310	0.086	0.015

What this does and does not show. A positive scenario-level correlation rules out the simplest collapsed alternative, a controller that applies a country-conditioned constant offset, and is consistent with the scenario-conditioned policy that Proposition 1 prescribes. It is *not* a direct verification of the closed-form shrinkage in part (ii), which makes a quantitative claim about the ratio $|\delta^\star|/|\Delta|$ that a marginal scatter cannot identify without controlling for $|\Delta|$. We therefore present this diagnostic as evidence of scenario-level conditioning, not as a tight numerical test of the oracle shrinkage factor.

A10 Component Decomposition

Table 3 is computed by replaying the logged per-scenario controller state of the seven headline runs. The replay reproduces every published per-country MIS to within 3×10^{-16} and the logged sparing probabilities to within 10^{-7}, so switching one stage off changes nothing else. This appendix gives the finer comparisons between the arms, an oracle calibration control, and the country win counts of every arm (Table 15).

Finer comparisons between the arms. Adding the prior to the plain consensus gives an MIS at least as low as Full DISCA on six of the seven backbones, so given the prior, plain consensus does as well. On Magistral the prior raises MIS, and DISCA without the prior is the best arm. Dropping the gated refinement lowers MIS slightly, by less than 0.005 on every backbone. The same-logit controls of App. A11 run without the prior, so they credit the full controller rather than a single stage.

Oracle calibration control. Part of what MIS rewards is calibration. Table 14 adds one control that divides the vanilla gap by a single temperature, T=8, picked from $\{1.5, 2, 3, 5, 8\}$ by its MIS, so it uses the human labels. This one temperature beats DISCA on four backbones. T=8 is the largest value tried, and MIS was still falling there on six backbones. A temperature picked per backbone

from the same grid beats DISCA on five. That oracle needs the labels, while DISCA reaches its gain without them.

Table 14: Oracle calibration control (macro MIS, lower is better), next to the vanilla and Full DISCA rows of Table 3. The temperature row uses the human labels to pick T.

Arm	Llama-70B	Magistral	Phi-4	Qwen3-VL	Qwen2.5-7B	Phi-3.5	Gemma-E2B
Vanilla	.849	.403	.454	.575	.453	.638	.495
Full DISCA	.668	.350	.346	.466	.362	.571	.478
Vanilla, T=8 (uses labels)	.720	.363	.381	.426	.359	.520	.425

Table 15: Countries (out of 20) with lower MIS than vanilla decoding, per arm and backbone. The last row uses the human labels to pick its temperature.

Arm	Llama-70B	Magistral	Phi-4	Qwen3-VL	Qwen2.5-7B	Phi-3.5	Gemma-E2B
Consensus	17	16	15	11	17	8	10
Consensus + prior	20	16	18	18	17	18	12
DISCA without prior	12	18	16	11	19	5	10
DISCA without refinement	20	16	18	20	17	17	14
Full DISCA	20	16	18	20	17	17	14
Vanilla, temperature T=8	20	14	18	20	19	20	19

A11 Same-Logit Controls and Transfer Benchmarks

This appendix gives the full tables behind Table 4, including the comparisons and prompt conditions summarised only briefly in §4.3.

Same-logit aggregation controls. For Qwen2.5-7B and Qwen2.5-14B we reuse, for every scenario, the saved base gap, the four persona gaps, and the decision temperature from the three-seed runs (App. A9.1), and recompute each aggregation rule on exactly these inputs. No comparator calls the model again. Table 16 reports the three-seed means. These runs are separate decodes from the headline sweep, so their vanilla and DISCA values differ from Table 2 (Qwen2.5-7B vanilla 0.453 and DISCA 0.362 there, 0.423 and 0.356 here). Comparisons hold within each table. On Qwen2.5-14B, DISCA is 0.008 below the vanilla of the same decode and has lower MIS than vanilla in 10 of 20 countries. DISCA has lower MIS than probability averaging, logit consensus, the median persona, and majority vote on both backbones. These rules aggregate the panel without the country prior, so the comparison credits the full controller rather than a single stage. Given the prior, plain consensus does as well: in the component decomposition, consensus with the prior is at least as low as Full DISCA on six of the seven backbones (App. A10).

Aggregation rules. Each rule maps the four persona gaps $\delta_1, \ldots, \delta_4$ of one scenario to a sparing probability, using the same decision temperature T_{dec} as DISCA. *Vanilla* uses the base gap, $\sigma(\delta_{\text{base}}/T_{\text{dec}})$. *Probability averaging* is the mean of the persona probabilities, $\frac{1}{4}\sum_i \sigma(\delta_i/T_{\text{dec}})$. *Logit consensus* applies the sigmoid to the mean gap, $\sigma(\bar{\delta}/T_{\text{dec}})$. *Median persona* applies it to the median gap. *Majority vote* returns 1 when at least three of the four gaps favour one side and 0 when at most one does; at a two-two tie it returns the vanilla probability. On NormAd and GlobalOpinionQA the same rules act on the K option scores: probability averaging takes the mean of the persona softmax distributions, and majority vote takes the option most personas rank first, with ties broken by the vanilla probabilities.

Table 16: Same-logit controls on MultiTP (macro MIS, lower is better, three-seed mean over 20 countries). These runs are separate decodes from the headline sweep, so the vanilla and DISCA values differ from Table 2. "Countries" counts the countries where the comparator has higher MIS than DISCA; a range gives the lowest and highest count over the three seeds.

	Qwen2.5-7B		Qwen2.5-14B	
Aggregation	MIS ↓	Countries	MIS ↓	Countries
Vanilla	0.4231	15	0.4032	10
Median persona	0.3866	16–17	0.4359	16
Probability averaging	0.3872	17–18	0.4342	16
Logit consensus	0.3877	18	0.4393	16
Majority vote (vanilla tie-break)	0.4307	16	0.4613	17
Full DISCA	**0.3563**		**0.3950**	

NormAd. NormAd [Rao et al., 2025] asks whether a described behaviour is acceptable in a given country, with three labels (yes, no, neutral). We use the country-context condition on 622 items over 17 countries with official WVS profiles, English personas and task language, and two to four seeded option orders mapped back to a canonical order. The controller never sees gold labels. Accuracy and macro-F1 are computed per country and averaged with equal weight. Here DISCA runs the K-way form of the controller: the option scores are centred to sum to zero, the persona mean is the anchor, and the two PT-IS passes and the gate act in that K-dimensional space; this setting uses no country prior, and its Prospect-Theory loss exponent is 1.0 rather than the 0.88 of the MultiTP controller. NormAd also offers two richer prompts from the same run, one that adds the country's values and one that adds a rule of thumb (lower blocks of Table 17). Under both, vanilla decoding is already more accurate than under the country-only prompt. With the values prompt no aggregation beats vanilla, and DISCA is 0.002 and 0.006 below it. With the rule of thumb, only probability averaging on Qwen2.5-14B beats vanilla, by 0.001. When the prompt already states the norm or the country's values, the model needs no cultural correction: the added context supplies what the panel would otherwise contribute, and a further correction can only move a decision that is already right.

Table 17: NormAd, equal-country mean, higher is better. All rows come from a single run with seed 42. Aggregation rules as in App. A11.

	Qwen2.5-7B		Qwen2.5-14B	
Method	Accuracy	Macro-F1	Accuracy	Macro-F1
Country-context condition				
Vanilla	0.5543	0.4613	0.5299	0.5014
Aggregate WVS profile prompt	0.5739	0.4846	0.5472	0.5064
Probability averaging	0.5756	0.4849	0.5497	0.5038
Majority vote	0.5735	0.4794	0.5463	0.4999
DISCA	0.5774	0.4879	0.5531	0.5077
Value and country condition				
Vanilla	0.6336	0.5343	0.6488	0.6030
Aggregate WVS profile prompt	0.6305	0.5196	0.6403	0.5816
Probability averaging	0.6315	0.5198	0.6441	0.5858
Majority vote	0.6315	0.5193	0.6417	0.5815
DISCA	0.6313	0.5201	0.6428	0.5850
Rule-of-thumb condition				
Vanilla	0.7476	0.6916	0.7834	0.7486
Aggregate WVS profile prompt	0.7073	0.6169	0.7828	0.7394
Probability averaging	0.7079	0.6183	0.7844	0.7417
Majority vote	0.7045	0.6134	0.7815	0.7376
DISCA	0.7079	0.6183	0.7827	0.7396

GlobalOpinionQA. GlobalOpinionQA [Durmus et al., 2024] pairs survey questions with the answer distribution of each country. We keep questions from the Pew Global Attitudes source that cover at least five of our target countries, remove all 353 questions taken from the World Values

Survey so that the personas cannot echo their own source data, and obtain 560 questions, 4,814 question-country pairs, and 15 countries. The metric is the base-2 Jensen-Shannon distance between the model's answer distribution and the country's survey distribution, averaged with equal country weight; lower is better. Table 18 reports the results. DISCA moves answers closer to the survey distributions than vanilla decoding, the aggregate WVS profile, and majority vote, but probability averaging over the same personas does better. We therefore read this benchmark as a boundary: a distributional target rewards averaging, while DISCA is built to return a single calibrated decision.

Table 18: GlobalOpinionQA, 560 questions, 15 countries (Jensen-Shannon distance, lower is better; the DISCA row averages the controller over seeds 42, 101 and 2026 on the same model outputs, App. A9.1; the other rows are deterministic).

Method	Qwen2.5-7B	Qwen2.5-14B
Vanilla	0.6130	0.6689
Aggregate WVS profile prompt	0.5931	0.6127
Probability averaging	**0.5692**	**0.5879**
Majority vote	0.6337	0.6322
DISCA	0.5838	0.6111

Persona order and country specificity. A fresh audit on the 310 United States scenarios with Qwen2.5-7B presents the four personas in canonical, reversed, and seeded random order. The largest absolute change in any final sparing probability is zero and no hard decision changes. A permutation placebo that reassigns country labels at random gives large lower-is-better p-values (about 0.75 to 1.00), so the MultiTP gain alone does not establish that each country's profile is specific to that country; the claim we make is that profile content is relevant, not that each profile is a faithful portrait.

A12 Baseline Implementation Details

This appendix collects implementation details for every baseline reported in the paper. §A12.1 covers the training-free baselines of Table 1 (vanilla decoding, WVS Profile Prompt, fixed logit offset). §A12.2 covers activation steering, and §A12.3 covers ARGS / controlled decoding. §A12.4 reports additional inference-time baselines run on the twenty-country Phi-4 grid (MC-Dropout, an attempted per-country temperature and margin oracle, and an AMCE-mixing oracle).

A12.1 Other training-free baselines (Table 1)

Vanilla decoding. The model is shown the native-language scenario with no persona prefix and no logit modification. We extract $\sigma(\delta(x)/T_{\text{dec}})$ and average across scenarios as in Eq. 9.

WVS Profile Prompt. A single country-conditioned system prompt is constructed by concatenating the ten WVS descriptors of Table 28 (population aggregate, no cohort split) into one paragraph in the country's native language; the model then answers each scenario under that prompt. This collapses the four-persona panel to a single point estimate, isolating the contribution of persona *diversity* on top of WVS grounding.

Fixed logit offset. A constant country-specific bias is added to the two decision-token logits, the same for every scenario in country c. The bias is read from the model and uses no human AMCE. The model answers country statements under the base prompt and the same A/B question frame as the scenarios, and the centred mean log-probabilities give $b_A(c) = \frac{1}{2}\big(\overline{\log p}(A) - \overline{\log p}(B)\big)$ and $b_B(c) = -b_A(c)$, which are added to z_A and z_B. The implementation averages over one statement per WVS profile dimension of the country. In the run reported here the statement set held a single sentence per country, and the resulting offsets are about 35 logits in size. The offset therefore pushes the sparing probability below 0.01 or above 0.99 in 95% of scenarios. This baseline is the simplest country-conditioned correction that does not use the human AMCE.

Persona-ensemble prompt. A demographic persona-conditioning baseline in the spirit of pluralistic persona prompting [Santurkar et al., 2023, Sorensen et al., 2024]: the model answers under six sociodemographic personas of the country (for example a young urban professional, an elderly religious person and a rural working-class adult), and the six sparing probabilities are averaged with fixed weights. No WVS statistics and no human AMCE enter the prompt. With long persona descriptions this baseline reaches a macro MIS of 0.364 on the Phi-4 grid, the strongest training-free baseline and still above DISCA's 0.346; with one-sentence descriptions it reaches 0.460, close to vanilla. Table 1 reports the long variant.

A12.2 Activation Steering

For each country, the steering vector is the difference between the mean last-token residual-stream activations of the positive and the negative prompts of five contrastive pairs, one per WVS dimension (religiosity, gender equality, tolerance of diversity, social trust and national pride). Each pair has the form "I am a person from [country] who is [pole]", with the country's WVS pole in the positive prompt and the opposite pole in the negative one. The vector is normalised to unit length and extracted from the same model it steers, so there is no cross-model transfer. At inference, $\alpha \cdot \mathbf{v}_{\text{steer}}$ is added to the residual stream at one layer on every forward pass. The Phi-4 run of Table 1 uses layer 32 of 40 and $\alpha = 1.5$ for every country.

Sensitivity analysis and fairness considerations. Activation steering depends on the extraction layer and position, which Arditi et al. [2024] select per model on validation data, and on the scaling coefficient α and the contrast pairs [Turner et al., 2023]. The reported run uses one setting, and a broader search over layers, coefficients and prompt designs could improve this baseline. At this setting it lowers macro MIS from 0.454 (vanilla) to 0.430, which remains above DISCA. A structural difference also remains: DISCA adapts the size of its correction to each scenario, whereas activation steering adds one fixed direction to every scenario.

A12.3 ARGS and controlled decoding (relation to our work)

ARGS [Khanov et al., 2024] and controlled decoding [Mudgal et al., 2024] typically assume trained reward models; a cross-cultural deployment would require a separate reward per target country. On binary forced-choice MultiTP, one can instead define rewards as closed-form functionals of persona logit gaps; a Prospect-Theory kernel aligned with Eq. 3 is then the natural analogue of importance-weighted PT-IS. We do not run the full ARGS decoding stack in our experiments; Table 20 isolates the kernel comparison on the shared DISCA pipeline.

A12.4 Additional inference-time baselines (twenty-country Phi-4 grid)

To isolate what within-country persona disagreement adds above simpler country-conditioned corrections, we compare DISCA against further inference-time alternatives on the same twenty-country Phi-4 grid. MC-Dropout is country-agnostic. MC-Dropout and DISCA are scored on the full evaluation pool used in the main results, so their macro means line up with Table 1 (MC-Dropout 0.403) and Table 2 (DISCA 0.346).

(6) MC-Dropout calibration. Following Kwon et al. [2026], who enable attention dropout at inference (rates 0.05 and 0.1), we enable every `torch.nn.Dropout*` module (rate 0.10) and average the A/B probabilities over eight stochastic forward passes per scenario. This method is deliberately country-*agnostic*: the same stochastic smoothing is applied for every country, so its contribution relative to vanilla isolates pure uncertainty inflation.

(7) Per-country temperature and margin oracle (attempted). We also attempted a per-country oracle. For each country it splits the 310 scenarios into 78 calibration and 232 test scenarios, and fits one scalar on the calibration set against the human AMCE: a temperature that divides the vanilla decision gap, or an additive margin (the bias term of Platt-style scaling [Guo et al., 2017]). The search grids, temperatures from 0.25 to 4 and margins from -4 to 4, did not reach the scale of the raw decision gaps this fit used. In 99.8% of scenarios the raw gap exceeds 20 logits in absolute value, so the fit landed on the edge of its grid for every country. Because this fit does not measure

what a calibration oracle can reach, we do not report it as a baseline. The calibration control we report is the label-picked global temperature of App. A10 (Table 14).

(8) AMCE-mixing oracle. DiffPO [Chen et al., 2025] trains a sentence-level refinement module; this oracle only borrows the idea of moving outputs toward a target at inference time. We mix the vanilla probability with a country-conditioned target $p_{\text{target}}(c, \text{cat})$ built directly from the country's *public* human AMCE ($p_{\text{aligned}} = (1-\alpha)\, p_{\text{van}} + \alpha\, p_{\text{target}}$), with $\alpha \in [0, 1]$ fit per country against the same AMCE. Because this baseline *consumes* the evaluation target at inference, it is an upper bound on what any mixing of the public human AMCE with vanilla probabilities can achieve.

Takeaway. DISCA never touches the human AMCE, while the AMCE-mixing oracle (8) uses it. The per-country results are in Table 19: DISCA wins **18 of 20** countries against MC-Dropout.

Table 19: Phi-4 (14B) per-country head-to-head: DISCA vs. MC-Dropout on the same 20 countries. Per-country MIS (↓, lower is better), both scored on all 310 scenarios. **Bold** cells mark the better method per country. The AMCE-mixing oracle is omitted: as a population-target replay it scores ≈ 0 MIS by construction (each country's prediction trivially copies the country's human AMCE) and is not a comparable inference-time baseline.

Country	DISCA (ours)	MC-Dropout
ARG	**0.389**	0.412
BGD	**0.368**	0.379
BRA	**0.299**	0.306
CHN	**0.214**	0.367
COL	0.438	**0.389**
DEU	**0.255**	0.376
ETH	**0.485**	0.526
GBR	**0.319**	0.407
IDN	**0.274**	0.452
IRN	**0.530**	0.538
JPN	**0.195**	0.371
KGZ	0.393	**0.361**
MEX	**0.420**	0.445
MMR	**0.357**	0.382
MYS	**0.328**	0.363
ROU	**0.385**	0.398
SRB	**0.385**	0.410
THA	**0.313**	0.342
USA	**0.243**	0.369
VNM	**0.336**	0.473
Macro	**0.346**	0.403
Wins (DISCA vs.)	-	18/20

A13 Hyperparameters and Sensitivity Sweeps

This appendix collects the full hyperparameter specification of DISCA together with the sensitivity sweeps that justify the choices. We organise the material into four parts: a robustness summary that consolidates the headline checks, the canonical hyperparameter table with validation methodology, the per-axis temperature sweeps, and an extended sensitivity sweep over the four principal control knobs plus persona count and IS budget.

A13.1 Robustness Summary

Table 20 consolidates the eight independent robustness checks referenced in §4.1. Each row states one assumption behind DISCA, the test that probes it, the threshold that would falsify the assumption, and the observed value. Per-axis sweeps and full per-country tables for each row appear in the subsections below and in App. A16.

Table 20: Robustness suite. Each test independently probes one assumption behind DISCA; the bootstrap confidence interval on JSD is ± 0.004. **Acronyms:** *CS-Clamp* = consensus-clamp deterministic shift (apply $\bar{\delta}$ directly, no IS); *ARGS-Unif* = ARGS-style reward-guided decoding [Khanov et al., 2024] with uniform reward over personas; *ARGS-WVS* = same with WVS-weighted reward; *ARGS-PT* = same with the Prospect-Theory kernel of Eq. 3.

Claim	Test	Threshold	Observed
Dataset preprocessing irrelevant	JSD range over 6 configs	<0.010	**0.0034**
T_{dec} insensitivity	JSD span (Sweep A)	<0.010	**0.0074**
Uniform T_c insensitivity	JSD span (Sweep B)	<0.010	**0.0073**
Per-category T_c insensitivity	JSD span (Sweep C)	<0.010	**0.0077**
PT-IS $>$ best deterministic shift	ΔJSD(CS-Clamp$\rightarrow$PT-IS)	>0	**0.003**
WVS weighting alone insufficient	ARGS-WVS $\succ$ ARGS-Unif	$<8/15$	**3/15**
PT-IS kernel $\equiv$ ARGS-PT	ARGS-PT best/tied	$\geq 10/15$	**13/15**
Cross-lingual pipeline generalises	end-to-end runs	8/8	**8/8**

A13.2 Default Hyperparameters and Validation

Table 21: DISCA hyperparameters. Prospect Theory parameters follow the estimates of Tversky and Kahneman [1992]; the other defaults were checked on the prototyping panel described below and then frozen.

Parameter	Symbol	Value
Persona agents	N	4
IS samples / pass	K_{half}	64 (total $2K_{\text{half}} = 128$)
Reliability scale	s	0.04
Perturbation std. dev.	σ	$\max\{\text{sd}(\delta_{1:N}), 0.3\}$
Country prior rate / warm-up / time scale	ω, t_0, τ	0.1 / 50 / 100
Cooperation weight	λ_{coop}	0.7
IS softmax temp.	η	0.5
ESS guard threshold	ρ_{eff}	0.1
PT curvature (gains and losses)	α	0.88
PT loss aversion	κ	2.25
Decision temp.	T_{dec}	0.5
Per-attribute logit temp. (Species / Gender / other)	T_c	4.0 / 3.5 / 1.5
T_c for a category without an entry	T_c	3.0
ESS anchor blend	α_{ess}	on (off: $\bar{\delta} + \delta^\star$ only)

Implementation. The dual-pass controller follows the equations in §3. Optional environment overrides can change K_{half}, reliability scale s, and whether the ESS anchor blend is active; such flags must be applied before controller initialisation when running ablations.

The non-Prospect-Theory hyperparameters (σ, λ_{coop}, η, T_{dec}, T_c) were checked on a small prototyping panel of three backbones and five countries and then frozen for every run reported in this paper. The T_c values reflect the empirical size of the decision gaps: Species and Gender produce systematically larger gaps and therefore get higher temperatures. Because the defaults were chosen on a small panel, we report broad sensitivity sweeps (App. A13, App. A16) to show that the conclusions do not rest on one tuned point.

A13.3 Temperature Sensitivity

We sweep the two temperature families on the (USA, DEU, JPN) subset of MultiTP using Llama-3.1-70B. Sweep A varies $T_{\text{dec}} \in \{0.10, 0.25, 0.50, 0.75, 1.0, 2.0\}$; Sweep B replaces the per-category temperatures by a single uniform $T_c \in \{0.5, 1.0, 1.5, 2.0, 3.0, 4.0, 6.0\}$; Sweep C varies the "Others" bucket $T_c \in \{0.75, 1.0, 1.5, 2.0, 3.0, 4.0\}$ while holding T_c[Species]$=4.0$, T_c[Gender]$=3.5$. All three sweeps yield JSD spans below 0.010 (0.0074, 0.0073, 0.0077 respec-

tively), and in every sweep the DISCA default (†) is strictly *more conservative* than the empirically best value (ΔJSD $\in \{-0.0061, -0.0021, -0.0037\}$).

Table 22: Temperature sensitivity. Defaults (†) are strictly worse than the best setting in every sweep, providing a conservative lower bound.

Sweep	Setting	JSD ↓	Pearson r ↑	ΔJSD
A: T_{dec}	0.10	0.0502	0.662	+0.0013
	0.25	0.0491	0.658	+0.0002
	0.50†	0.0489	0.630	-
	0.75	0.0476	0.636	−0.0013
	1.00	0.0466	0.641	−0.0023
	2.00 (best)	0.0428	0.627	−0.0061
	JSD span		**0.0074**	
B: Uniform T_c	0.5	0.0521	0.642	+0.0052
	1.0	0.0492	0.673	+0.0023
	1.5†	0.0469	0.700	-
	2.0	0.0461	0.707	−0.0008
	3.0	0.0454	0.707	−0.0015
	4.0	0.0451	0.701	−0.0018
	6.0 (best)	0.0448	0.684	−0.0021
	JSD span		**0.0073**	
C: Others T_c	0.75	0.0526	0.566	+0.0039
	1.00	0.0512	0.589	+0.0025
	1.50†	0.0487	0.634	-
	2.00	0.0475	0.660	−0.0012
	3.00	0.0458	0.693	−0.0029
	4.00 (best)	0.0450	0.706	−0.0037
	JSD span		**0.0077**	

The monotonic JSD reduction with larger T_{dec} reflects the role of T_{dec} in undoing RLHF logit compression: higher values widen the effective token-probability landscape so that IS perturbations exert larger directional influence. However, r peaks near T_{dec}=1.0 and declines at 2.0, indicating a trade-off between distributional alignment (JSD) and rank-order alignment (r); we keep T_{dec}=0.5 as a deliberately conservative operating point.

A13.4 Extended Hyperparameter Sensitivity

We probe the sensitivity of DISCA to the four principal hyperparameters: s (reliability gate scale), λ_{coop} (individual-vs-consensus weight), σ (IS proposal floor), and T_c (uniform logit-temperature scaling). For each axis we sweep five values on a three-country panel (USA, VNM, DEU) and hold the other three axes at their defaults; the s axis directly probes how the choice of reliability scale in Eq. 5 affects downstream MIS.

Table 23: Hyperparameter sensitivity ranges on the three-country panel (Phi-4, 310 scenarios per country). Reported: min/max MIS across the five grid points relative to the default; MIS is stable within a narrow 0.0037 to 0.0090 span on every axis, so the gains in Table 2 are not the product of a particular hyperparameter choice.

Axis	Default	Grid	Min MIS	Max MIS	Δ
s	0.04	$\{0.01, 0.02, 0.04, 0.08, 0.16\}$	0.4091	0.4164	0.0073
λ_{coop}	0.70	$\{0.30, 0.50, 0.70, 0.85, 0.95\}$	0.4085	0.4173	0.0088
σ	0.30	$\{0.10, 0.20, 0.30, 0.45, 0.60\}$	0.4084	0.4122	0.0037
T_c scale	1.0×	$\{0.33\times, 0.67\times, 1.0\times, 1.33\times, 1.67\times\}$	0.4060	0.4149	0.0090

Why these four axes. s controls the steepness of the reliability gate (Eq. 5); λ_{coop} trades off individual-persona PT utility against the consensus utility; σ floors the IS proposal scale so the $N{=}4$ empirical std cannot collapse to zero; and T_c scales the whole family of per-category decision temperatures (preserving relative balance), a global sharpening or flattening knob on the decision logits. Per-category temperatures are stress-tested separately in the temperature sensitivity sweep above.

Persona count, IS budget, and per-category vs. global T_c. We additionally swept three operational knobs on the same USA/VNM/DEU panel. **Persona count** ($N \in \{2, 3, 4, 5, 6\}$, Table 24): $N{=}4$ is optimal at macro MIS 0.4051, with degradation at both smaller and larger panels (0.4252 at $N{=}2$, 0.4179 at $N{=}6$), supporting the four-persona design. **IS budget** ($K_{\text{half}} \in \{8, 16, 32, 64, 128, 192\}$, Table 25): performance is flat across an order of magnitude (best 0.4092 at $K{=}16$; 0.4103 at $K{=}64$), indicating the default $K_{\text{half}}{=}64$ is in the stable region rather than overtuned. **Global vs. per-category** T_c (Table 26): the per-category default outperforms *every* global schedule we tested (0.4098 vs. all seven global values in $\{1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$, which land in 0.4119 to 0.4138). The advantage is small in absolute terms but consistent.

Table 24: Persona-count sweep on the USA/VNM/DEU panel (Phi-4, 310 scenarios per country). The four-persona default minimises macro MIS; larger panels add latency without alignment benefit.

N	Macro MIS ↓	Std (across countries)	Flip rate	sec/scenario
2	0.4252	0.1045	0.275	0.072
3	0.4194	0.0945	0.273	0.080
4	**0.4051**	0.0829	0.276	0.090
5	0.4107	0.0745	0.276	0.097
6	0.4179	0.0919	0.245	0.105

Table 25: IS-budget sweep (K_{half} samples per pass; total per scenario is $K_{\text{total}}{=}2K_{\text{half}}$). Macro MIS is flat across an order of magnitude. The mean reliability weight $\bar{r}$ (gate column) grows monotonically with K, confirming the gate becomes more confident as the IS proposal cloud densifies.

K_{half}	K_{total}	Macro MIS ↓	Flip rate	$\bar{r}$ (gate)	sec/scen
8	16	0.4114	0.284	0.684	0.097
16	32	0.4092	0.271	0.736	0.092
32	64	0.4104	0.278	0.766	0.092
64	128	0.4103	0.280	0.793	0.091
128	256	0.4147	0.274	0.814	0.093
192	384	0.4126	0.273	0.827	0.091

Table 26: Global vs. per-category T_c on the USA/VNM/DEU panel. The per-category default (Species 4.0, Gender 3.5, other attributes 1.5) outperforms every global value we tested. The gap is small (≤ 0.004) but uniformly directional, supporting category-specific decision temperatures over a single shared scalar.

T_c schedule	Macro MIS ↓	Std	Flip rate
per-category default (4.0 / 3.5 / 1.5)	**0.4098**	0.0881	0.278
global $T_c = 1.00$	0.4120	0.0921	0.292
global $T_c = 1.50$	0.4119	0.0838	0.304
global $T_c = 2.00$	0.4138	0.0778	0.301
global $T_c = 2.50$	0.4134	0.0774	0.296
global $T_c = 3.00$	0.4127	0.0749	0.303
global $T_c = 3.50$	0.4138	0.0689	0.306
global $T_c = 4.00$	0.4128	0.0734	0.302

Per-persona utility floor (minority safeguard). The released controller optionally applies a per-persona utility floor that limits how far any single persona's post-correction utility can fall below

vanilla. Sweeping $f \in \{0.0, 0.5, 1.0, 2.0\}$ on the same USA/VNM/DEU panel (Phi-4, 310 scenarios per country) keeps macro MIS in $[0.4046, 0.4157]$, well inside the macro-MIS bootstrap noise floor, so the floor is a stability knob rather than a load-bearing hyperparameter.

Table 27: Per-persona utility floor sweep (USA / VNM / DEU). Macro MIS is flat across two orders of magnitude of the floor parameter.

Floor f	Macro MIS ↓	Std (across countries)
0.0 (off)	0.4083	0.0717
0.5	**0.4046**	0.0696
1.0	0.4066	0.0751
2.0	0.4157	0.0684

A14 Persona Construction Details

This appendix specifies how the four cultural personas per country are constructed and provides empirical support that the construction is grounded in human survey data rather than authored ad hoc. §A14.1 reproduces two personas verbatim from `generate_wvs_persona()`; §A14.2 documents the WVS Wave 7 processing pipeline; and §A14.3 links the ten WVS cultural dimensions to the six MultiTP moral attributes both thematically and empirically.

A14.1 Persona Prompts

Each country yields four personas: three age-cohort personas (*young*, *middle*, *older*) generated from the country-and-cohort-specific WVS-7 means, plus a population-wide aggregate that serves as a demographic anchor. Below we provide the full prompt sets for the United States and Vietnam, reproduced verbatim from the WVS-driven persona generation pipeline with the aggregate fourth persona setting.

United States (all four personas, verbatim)

Young cohort.

> *You are a young adult from the United States, in your 20s and early 30s. Your worldview is shaped by the cultural values prevalent in your community. On matters of faith you are moderately religious. On raising children you are firmly oriented toward independence and imagination. On contested moral choices you are morally conservative on contested issues. In your dealings with strangers you have a guarded attitude toward strangers. Civically you are an active political participant who signs petitions, joins boycotts and takes part in lawful demonstrations. You are moderately proud of your country. Overall you are rather happy with your life. On the role of women in society you are moderately egalitarian on gender roles. In what you prioritise in life you are leaning materialist, prioritising economic and physical security. Toward people unlike yourself you are highly tolerant of outgroups such as immigrants, minorities and people with different lifestyles. When you face a moral dilemma, you weigh the choices through this set of values and answer in a way that is consistent with the worldview above.*

Middle-aged cohort.

> *You are a middle-aged adult from the United States, in your 40s or 50s. Your worldview is shaped by the cultural values prevalent in your community. On matters of faith you are moderately religious. On raising children you are firmly oriented toward independence and imagination. On contested moral choices you are strictly opposed to such contested moral acts. In your dealings with strangers you have a guarded attitude toward strangers. Civically you are an active political participant who signs petitions, joins boycotts and takes part in lawful demonstrations. You are moderately proud of your country. Overall you are rather happy with your life. On the role of women in society you are moderately egalitarian on gender roles. In what you prioritise in life you are leaning materialist, prioritising economic and physical security. Toward people unlike yourself you are highly tolerant of outgroups such as immigrants, minorities and people with different lifestyles.*

When you face a moral dilemma, you weigh the choices through this set of values and answer in a way that is consistent with the worldview above.

Older cohort.

You are a senior citizen from the United States, over 60 years old. Your worldview is shaped by the cultural values prevalent in your community. On matters of faith you are moderately religious. On raising children you are firmly oriented toward independence and imagination. On contested moral choices you are strictly opposed to such contested moral acts. In your dealings with strangers you have a guarded attitude toward strangers. Civically you are an active political participant who signs petitions, joins boycotts and takes part in lawful demonstrations. You are intensely proud of your country. Overall you are rather happy with your life. On the role of women in society you are moderately egalitarian on gender roles. In what you prioritise in life you are leaning materialist, prioritising economic and physical security. Toward people unlike yourself you are highly tolerant of outgroups such as immigrants, minorities and people with different lifestyles. When you face a moral dilemma, you weigh the choices through this set of values and answer in a way that is consistent with the worldview above.

Aggregate cohort (all ages).

You are a adult citizen from the United States. Your worldview is shaped by the cultural values prevalent in your community. On matters of faith you are moderately religious. On raising children you are firmly oriented toward independence and imagination. On contested moral choices you are strictly opposed to such contested moral acts. In your dealings with strangers you have a guarded attitude toward strangers. Civically you are an active political participant who signs petitions, joins boycotts and takes part in lawful demonstrations. You are moderately proud of your country. Overall you are rather happy with your life. On the role of women in society you are moderately egalitarian on gender roles. In what you prioritise in life you are leaning materialist, prioritising economic and physical security. Toward people unlike yourself you are highly tolerant of outgroups such as immigrants, minorities and people with different lifestyles. When you face a moral dilemma, you weigh the choices through this set of values and answer in a way that is consistent with the worldview above.

Vietnam (all four personas, Vietnamese, verbatim)

Young cohort.

Bạn là một thanh niên đến từ Việt Nam, ở độ tuổi 20 đến đầu 30. Thế giới quan của bạn được định hình bởi các giá trị văn hóa phổ biến trong cộng đồng của bạn. Về vấn đề đức tin, bạn khá thế tục. Trong việc nuôi dạy con cái, bạn nghiêng về sự vâng lời và đức tin tôn giáo. Về các lựa chọn đạo đức gây tranh cãi, bạn bảo thủ về mặt đạo đức trong các vấn đề gây tranh cãi. Trong giao tiếp với người lạ, bạn có sự ngờ vực sâu sắc đối với người khác. Về mặt công dân, bạn là một người tham gia chính trị thụ động. Bạn rất tự hào về đất nước của bạn. Nhìn chung, bạn rất hài lòng với cuộc sống của bạn. Về vai trò của phụ nữ trong xã hội, bạn khá truyền thống về vai trò giới. Về những gì bạn ưu tiên trong cuộc sống, bạn có xu hướng hậu vật chất. Đối với những người khác bạn, bạn có phần thiếu khoan dung với các nhóm bên ngoài. Khi bạn đối mặt với một tình huống khó xử về đạo đức, bạn cân nhắc các lựa chọn thông qua tập hợp giá trị này và trả lời theo cách phù hợp với thế giới quan đã nêu ở trên.

Middle-aged cohort.

Bạn là một người trung niên đến từ Việt Nam, ở độ tuổi 40 hoặc 50. Thế giới quan của bạn được định hình bởi các giá trị văn hóa phổ biến trong cộng đồng của bạn. Về vấn đề đức tin, bạn khá thế tục. Trong việc nuôi dạy con cái, bạn nghiêng về sự vâng lời và đức tin tôn giáo. Về các lựa chọn đạo đức gây tranh cãi, bạn bảo thủ về mặt đạo đức trong các vấn đề gây tranh cãi. Trong giao tiếp với người lạ, bạn có thái độ dè dặt với người lạ. Về mặt công dân, bạn là một người thỉnh thoảng tham gia chính trị. Bạn rất tự hào về đất nước của bạn. Nhìn chung, bạn rất hài lòng với cuộc sống của bạn. Về vai trò của phụ nữ trong xã hội, bạn khá truyền thống về vai trò giới. Về những gì bạn ưu tiên trong cuộc sống, bạn có xu hướng hậu vật chất. Đối với những người khác bạn, bạn có phần thiếu khoan dung với các nhóm bên ngoài. Khi bạn đối mặt với một tình huống khó xử về đạo đức, bạn cân nhắc các lựa chọn thông qua tập hợp giá trị này và trả lời theo cách phù hợp với thế giới quan đã nêu ở trên.

Older cohort.

> *Bạn là một người cao tuổi đến từ Việt Nam, trên 60 tuổi. Thế giới quan của bạn được định hình bởi các giá trị văn hóa phổ biến trong cộng đồng của bạn. Về vấn đề đức tin, bạn có niềm tin tôn giáo ở mức trung bình. Trong việc nuôi dạy con cái, bạn nghiêng về sự vâng lời và đức tin tôn giáo. Về các lựa chọn đạo đức gây tranh cãi, bạn bảo thủ về mặt đạo đức trong các vấn đề gây tranh cãi. Trong giao tiếp với người lạ, bạn có thái độ dè dặt với người lạ. Về mặt công dân, bạn là một người thỉnh thoảng tham gia chính trị. Bạn rất tự hào về đất nước của bạn. Nhìn chung, bạn rất hài lòng với cuộc sống của bạn. Về vai trò của phụ nữ trong xã hội, bạn khá truyền thống về vai trò giới. Về những gì bạn ưu tiên trong cuộc sống, bạn có xu hướng hậu vật chất. Đối với những người khác bạn, bạn rất thiếu khoan dung với các nhóm bên ngoài. Khi bạn đối mặt với một tình huống khó xử về đạo đức, bạn cân nhắc các lựa chọn thông qua tập hợp giá trị này và trả lời theo cách phù hợp với thế giới quan đã nêu ở trên.*

Aggregate cohort (all ages).

> *Bạn là một công dân trưởng thành đến từ Việt Nam. Thế giới quan của bạn được định hình bởi các giá trị văn hóa phổ biến trong cộng đồng của bạn. Về vấn đề đức tin, bạn khá thế tục. Trong việc nuôi dạy con cái, bạn nghiêng về sự vâng lời và đức tin tôn giáo. Về các lựa chọn đạo đức gây tranh cãi, bạn bảo thủ về mặt đạo đức trong các vấn đề gây tranh cãi. Trong giao tiếp với người lạ, bạn có thái độ dè dặt với người lạ. Về mặt công dân, bạn là một người thỉnh thoảng tham gia chính trị. Bạn rất tự hào về đất nước của bạn. Nhìn chung, bạn rất hài lòng với cuộc sống của bạn. Về vai trò của phụ nữ trong xã hội, bạn khá truyền thống về vai trò giới. Về những gì bạn ưu tiên trong cuộc sống, bạn có xu hướng hậu vật chất. Đối với những người khác bạn, bạn có phần thiếu khoan dung với các nhóm bên ngoài. Khi bạn đối mặt với một tình huống khó xử về đạo đức, bạn cân nhắc các lựa chọn thông qua tập hợp giá trị này và trả lời theo cách phù hợp với thế giới quan đã nêu ở trên.*

Aggregate (fourth) persona. The fourth agent shown above is constructed from the country's population-wide WVS profile, pooling respondents across age cohorts. It serves as a demographic anchor: by suppressing cohort-specific noise it ensures that the ensemble's consensus target $\bar{\delta}$ reflects the broadest available empirical signal rather than a single generational viewpoint.

Faithfulness disclaimer. The descriptors above are read directly from the country's WVS-7 means through deterministic quartile cuts (Table 28); they are not hand-selected to match cultural stereotypes. As a consequence, individual cohorts can present empirically grounded but locally counter-intuitive combinations (for example, a U.S. young-adult that is *leaning materialist* and *morally conservative on contested issues*), reflecting actual WVS-7 distributions rather than authorial bias.

A14.2 WVS Data Processing Pipeline

We process WVS Wave 7 individual-level microdata following the ten-variable cultural-value scheme of Greco et al. [2026]. The pipeline is fully deterministic: given the inverted WVS-7 CSV and the country code, it returns the four personas used at inference time without any tunable hyper-parameters.

1. **Variable extraction:** For each respondent, we extract 10 value dimensions from WVS items (Table 28). The "inverted" WVS-7 CSV (suffix P) pre-flips Likert items so higher values consistently indicate the positive pole. Items without the P suffix (Q152, Q153, Q177 to Q182) use original WVS coding.
2. **Age cohort assignment:** Using birth year (Q261) and survey year (A_YEAR), we compute age and assign to: *young* ($<$36), *middle* (36 to 55), *older* ($>$55). Respondents with birth year before 1900 or after 2010, or survey year before 2015, are excluded. Negative WVS codes ($-1, -2, -4, -5$: refusal / don't know) are dropped; zero is retained for binary 0/1 items.
3. **Cohort aggregation:** For each country and age cohort, each dimension's score is the unweighted mean of all valid (respondent, item) values associated with that dimension - i.e., for multi-item dimensions (e.g., Q177 to Q182 for moral acceptability) every respondent contributes one value per item to a single pooled mean. For single-item dimensions this reduces to the standard across-respondent mean.

4. **Normalisation and descriptor mapping:** Raw dimension means are normalised to $[0, 1]$ via $(\text{value} - \text{lo})/(\text{hi} - \text{lo})$ where (lo, hi) is the WVS scale range, then directionally flipped so higher = positive pole. Four-level descriptors are assigned by quartile cuts: ≥ 0.75 (strong positive), ≥ 0.50, ≥ 0.25, < 0.25 (strong negative). See Table 28.

5. **Fallback:** For countries with insufficient WVS coverage (e.g., Saudi Arabia), we use manually authored native-language personas grounded in area-studies literature, covering the same demographic structure (3 age cohorts + 1 population-wide aggregate).

Table 28: The 10 WVS cultural-value dimensions used for persona generation, following Greco et al. [2026]. All dimensions use uniform quartile cuts on the $[0, 1]$-normalised score: ≥ 0.75 (descriptor 1, strongest positive pole), ≥ 0.50 (2), ≥ 0.25 (3), < 0.25 (4, strongest negative pole). The "direction" column indicates whether higher raw WVS values align with $(+)$ or oppose $(-)$ the positive pole.

Dimension	WVS items	Range	Dir.	Descriptors ($\geq$0.75 / $\geq$0.50 / $\geq$0.25 / $<$0.25)
Religiosity	Q6P	1–4	$+$	deeply / moderately religious / somewhat / highly secular
Child-rearing	Q17P	0–1	$-$	independence-imagination $\leftrightarrow$ obedience-faith
Moral accept.	Q177–Q182	1–10	$+$	very permissive $\leftrightarrow$ strictly opposed
Social trust	Q57P	1–2	$+$	very high trust $\leftrightarrow$ deep distrust
Polit. particip.	Q199P, Q200P	1–3	$+$	active participant $\leftrightarrow$ non-participant
National pride	Q254P	1–4	$+$	very proud $\leftrightarrow$ not proud
Happiness	Q46P	1–4	$+$	very happy $\leftrightarrow$ not happy
Gender equality	Q29P, Q30P, Q31P, Q33P	1–4	$-$	strongly egalitarian $\leftrightarrow$ traditional roles
Materialism	Q152, Q153	1–3	$+$	post-materialist $\leftrightarrow$ materialist
Tolerance	Q19P–Q23P	0–1	$-$	very tolerant $\leftrightarrow$ intolerant

A14.3 WVS-to-Trolley Dimension Linkage

The linkage between the ten WVS cultural dimensions and the six MultiTP moral dimensions operates through two complementary mechanisms.

Direct thematic overlap. WVS *gender equality* maps onto the Gender dimension: the preference for sparing women is stronger in countries where women have better health and survival prospects [Awad et al., 2018]; *religiosity* correlates with Species through religious anthropocentrism and human exceptionalism; and *moral acceptability* relates to tolerance for trade-offs along the Fitness and Social Value dimensions.

Indirect modulation. Dimensions such as *social trust*, *child-rearing values*, and *materialism orientation* reshape the moral *frame* the persona adopts: post-materialist personas favour utilitarian calculi, while obedience-oriented child-rearing values up-weight hierarchical social roles (affecting Age and Social Value). Importantly, DISCA does not require an exact causal mapping between WVS and MultiTP dimensions; the ten-dimensional persona ensemble produces diversity in logit-space gaps, and the controller turns these gaps into one correction.

Empirical validation of the WVS$\rightarrow$AMCE pathway. Table 29 regresses each of the six MultiTP human AMCE dimensions on the 10 WVS cultural features across the 20-country panel (standardised OLS). The 10-feature WVS profile explains $R^2 \in [0.55, 0.69]$ of human AMCE variance, with $|\beta| > 0.3$ in 31 of 60 cells. This is direct evidence that WVS-grounded persona construction is not arbitrary: the cultural axes our personas inherit correlate strongly with the moral preferences DISCA must reproduce.

Table 29: WVS features as predictors of human AMCE dimensions. OLS with standardised predictors. Coefficients are standardised (β); bold = $|\beta| > 0.3$. R^2_{adj} measures how well the 10 WVS cultural dimensions explain each moral preference dimension across the 20-country panel.

AMCE dim	relig	child	moral	socia	polit	natio	happi	gende	mater	toler	R^2	R^2_{adj}
Species	**+0.47**	+0.12	**-0.65**	**+0.60**	+0.08	+0.16	-0.02	**-0.74**	-0.30	**+0.40**	0.57	0.09
Gender	**+0.69**	**+0.37**	-0.18	**+0.41**	**-0.44**	**-0.39**	**+0.71**	**-0.81**	-0.20	+0.06	0.66	0.29
Age	**-0.80**	-0.09	**+0.61**	**-0.74**	+0.17	+0.26	-0.27	**+0.62**	**+0.49**	**-0.37**	0.55	0.05
Fitness	+0.28	-0.23	-0.16	-0.12	-0.17	**+0.43**	+0.09	**-1.46**	-0.21	**+0.53**	0.60	0.15
SocialValue	-0.27	-0.30	**+0.40**	**-0.40**	**-0.37**	**+0.58**	+0.17	**-0.31**	-0.12	-0.15	0.69	0.35
Utilitarianism	**+1.11**	**+0.39**	+0.07	**+0.59**	-0.21	-0.01	+0.13	**-0.76**	**+0.42**	+0.25	0.66	0.29

Table 30 closes the loop with a causal leave-one-WVS-dim-out probe: for each WVS dimension d, we re-run DISCA with d removed from every persona and measure the change in per-MultiTP-dim absolute error (positive cell = error rises when the dimension is removed = that dimension is load-bearing for the corresponding moral attribute). The three largest positive couplings are *moral acceptability* $\rightarrow$ Social Value ($+3.97$ pp), *social trust* $\rightarrow$ Social Value ($+2.81$ pp), and *child-rearing values* $\rightarrow$ Utilitarianism ($+2.41$ pp). Cells with the top-3 largest positive couplings per row are bolded; rows with fewer than three positive cells are bolded only on those positive cells. Negative cells are also informative: *tolerance of diversity* $\rightarrow$ Social Value (-7.76) and *gender equality* $\rightarrow$ Gender (-3.42) indicate that, for those moral attributes, the corresponding WVS descriptor biases the persona ensemble *away* from the human target, so removing the descriptor improves accuracy. The loss-averse utility (Eq. 3) penalises candidate perturbations that worsen any persona's alignment (κ=2.25 for losses against gains). This penalty acts only on the refinement $\delta^\star$, so we do not claim that the controller filters out a biased descriptor; the dropout results above show which descriptors the panel depends on.

Table 30: WVS-dim $\times$ MultiTP-dim causal impact matrix. Cells are the mean increase in per-dim AMCE error (pp) when the WVS dimension is dropped from every persona, macro-averaged across the country panel. **Bold** cells mark the top-3 largest positive couplings per row (load-bearing WVS $\rightarrow$ MultiTP links).

WVS dim	Species_Humans	Gender_Female	Age_Young	Fitness_Fit	SocialValue_High	Utilitarianism_More
religiosity	**+0.66**	**+1.65**	-1.66	+0.58	**+1.02**	-0.37
child rearing	-1.82	**+1.29**	+0.86	**+1.30**	-0.33	**+2.41**
moral acceptability	**+1.13**	-0.02	-1.56	-0.34	**+3.97**	-1.69
social trust	+0.11	**+1.43**	-0.85	**+0.51**	**+2.81**	-1.88
political participation	-1.24	-0.82	-0.63	-0.36	-0.16	**+0.86**
national pride	-0.37	-0.10	-0.28	**+0.52**	**+1.19**	**+1.21**
happiness	-0.23	-0.88	**+0.13**	-1.69	-0.11	**+1.72**
gender equality	**+0.41**	-3.42	-2.90	-1.11	-1.87	**+0.62**
materialism orientation	**+1.35**	-1.26	**+0.90**	+0.04	-0.36	**+2.15**
tolerance diversity	-0.76	-3.52	**+1.95**	-0.59	-7.76	**+1.54**

A14.4 Macro-level WVS dimension importance

The per-MultiTP-dim impact matrix above (Table 30) tells us *which* moral attribute each WVS descriptor influences. A coarser but operationally useful question is which WVS dimensions are load-bearing for macro alignment, taken as a whole. We answer this with a leave-one-WVS-dim-out probe: for each of the ten descriptors, we re-run DISCA with that descriptor removed from every persona and record the change in macro MIS, averaged across USA, JPN, and VNM as a representative panel.

The result is a concentrated importance profile: religiosity, gender equality, and moral acceptability each raise MIS by more than 0.03 when removed; national pride and happiness barely register (< 0.01). The implication for persona engineering is that the ten-dimensional WVS profile can be compressed substantially while preserving most of the steering signal.

Table 31: WVS dimension dropout (3-country average; higher ΔMIS means more load-bearing). Dimension names follow the canonical WVS-7 list of Table 28 and Table 30.

Dropped WVS dimension	ΔMIS (avg)
religiosity	+0.046
gender equality	+0.041
moral acceptability	+0.034
social trust	+0.022
child rearing	+0.019
political participation	+0.017
materialism orientation	+0.013
tolerance diversity	+0.011
national pride	+0.007
happiness	+0.005

A14.5 Wrong-country profile test

Setup. This test asks whether DISCA uses the content of the WVS profile or only the presence of four personas. For each target country we keep the MultiTP scenarios, the prompt language and the country named in the persona header. We replace only the ten-dimensional WVS profile that describes the four personas with the profile of another country. Two one-to-one reassignments are used. The *random* condition draws a seeded derangement, so no country keeps its own profile. The *most distant* condition takes the derangement that maximises the total distance between the normalised WVS profiles. We run Qwen2.5-7B and Qwen2.5-14B with seed 42 on all 20 countries, 310 scenarios each. The matched condition is the seed-42 run of the three-seed set (App. A9.1), and vanilla decoding is identical across the three conditions.

Table 32: Wrong-country profile test (macro MIS over 20 countries, lower is better, seed 42). Change: relative change of macro MIS against the matched profile. Matched better: countries where the matched profile has lower MIS.

Backbone	Profile given to the personas	MIS ↓	Change (%)	Matched better
Qwen2.5-7B	Matched country	**.356**		
	Random wrong country	.357	+0.28	11/20
	Most distant wrong country	.359	+0.77	11/20
Qwen2.5-14B	Matched country	**.395**		
	Random wrong country	.412	+4.45	12/20
	Most distant wrong country	.416	+5.40	10/20

Result. All four wrong-profile conditions raise macro MIS, so the profile content matters. The effect is small on 7B and larger on 14B. It is not uniform across countries: the matched profile is better in only 10 to 12 of the 20 countries. The test shows that DISCA depends on which profile it is given. It does not show that each persona is a faithful portrait of its cohort (Table 36).

A15 AMCE Estimation Details

Uniform treatment of all dimensions. All six dimensions-Species, Gender, Age, Fitness, Social Value, and Utilitarianism-are computed identically. For each dimension, we compute the empirical mean of $p_{\text{spare}}(\mathbf{x})$ across all scenarios in that dimension (Eq. 9). For the five binary dimensions, each scenario presents a forced choice between two character groups differing on exactly one attribute. For Utilitarianism, scenarios vary the number of lives on each side; the binary indicator is whether the larger group ("More") or the smaller group ("Less") is spared. This uniform mean-based estimation is consistent with the `compute_ACME` method in the MultiTP codebase [Jin et al., 2025], which fits a no-intercept linear regression with a binary group indicator-mathematically equivalent to taking the mean of the saving probability for the preferred group.

Human AMCE scale. The MultiTP ground-truth AMCEs in the MultiTP-released country-specific AMCE table is on a $[-1, 1]$ scale. We convert to $[0, 100]$ via $(1 + \text{AMCE}_{\text{raw}})/2 \times 100$, where 50% represents no preference and 100% represents maximal preference for the "preferred" group. Model AMCEs are computed on the same scale.

Two reporting scales (resolving an apparent contradiction). The PCA visualisation in Figure 3 reports ℓ_2 distances on the $[0, 100]$ percentage-point scale, measured in the two-component PCA plane ($83.3 \rightarrow 64.6$ for Llama-3.3-70B; the full six-dimensional distance is $84.9 \rightarrow 66.8$), while every numerical table in the main text and appendix (Tables 1, 2, 7, etc.) reports MIS on the $[0, 1]$ proportional scale (i.e., directly in p_{spare} space, with values in $[0.2, 1.0]$). The two scales differ by a factor of 100. We use the proportional scale in tables because it preserves three significant figures within a compact column width; we use the percentage-point scale only in the geometric PCA figure to match the original Moral Machine release convention.

Consistency. Under the balanced randomisation design of the Moral Machine [Awad et al., 2018], marginal averages are unbiased estimators of the AMCE [Hainmueller et al., 2014], making our estimation valid and reproducible.

Cross-lingual A/B token elicitation. Eq. 9 requires a clean A vs. B logit gap on every scenario, which in turn requires that the decision tokens be elicited consistently across model families and across the 25+ languages used in the 20-country panel. Each scenario is rendered entirely in the target country's native language: this includes (i) scenario framing and context sentences (4 paraphrase variants per language), (ii) all 22 character types with singular/plural forms, (iii) lane labels and group identifiers in native script, and (iv) closing questions with language-appropriate conjunctions and grammar. The prompt frame wraps the scenario in a native-language instruction that explicitly requests an English answer token, following the pattern: "[moral dilemma preamble in native language] \n {scenario} \n [instruction to answer A or B in English]". This keeps moral reasoning in the culturally native linguistic frame while maintaining a single decision interface across the panel.

Token verification. Token IDs are verified per model at initialisation by encoding the uppercase strings `"A"` and `"B"` with `add_special_tokens=False` and asserting single-token mappings: Qwen2.5: A=22397, B=11572; Llama-3.1: A=31266, B=9268. Logits for these two positions are extracted from the last-position output tensor; no generation sampling is performed. The model uses its built-in `chat_template` for correct role formatting across model families.

A16 Dataset Preprocessing and Sensitivity

Pipeline. Each MultiTP CSV is processed in five steps: (i) deduplication keeps only the canonical paraphrase (`which_paraphrase = 0`); (ii) the Utilitarianism *quality filter* drops scenarios where the two groups have identical size *and* all characters belong to the *quality-attribute* role set {PREGNANT, WOMAN, LARGEWOMAN}, which would otherwise inject a non-numerosity signal into a dimension whose AMCE is by construction the count-difference effect; (iii) per-category capping retains at most 80 scenarios per dimension; (iv) under-represented categories are oversampled with replacement to a minimum of 36 scenarios; and (v) the resulting pool is shuffled with a fixed seed (42).

Reproducible left/right assignment. When the MultiTP `paraphrase_choice` field does not unambiguously determine which group is rendered on the left vs. right, we deterministically fall back to a SHA-256 hash of the tuple (sub_1, sub_2, g_1, g_2) modulo 2, so the same scenario receives the same ordering across runs and machines. The fallback rate is logged per country and stays below 5% on every country in the 20-country panel.

Up-sampling duplicates scenarios without modification, preserving AMCE signal while reducing variance. Category capping prevents dimension dominance. Side balancing enables effective debiasing. To evaluate whether these choices alter the effective evaluation distribution, we compare *six* preprocessing configurations on 10 representative countries (USA, DEU, CHN, JPN, BRA, VNM, GBR, KOR, RUS, NGA) using Llama-3.1-70B.

Table 33: Per-country dataset sensitivity. JSD range = 0.0034 (aggregate) and < 0.010 in 9/10 countries; the only exception is RUS at 0.013, driven by augmentation sensitivity in the rare Species/Utilitarianism categories. D0 is retained as default.

Country	D0-Default	D1-NoAug	D2-Cap40	D3-Cap120	D4-NoFlip	D5-Strict
USA	0.0655	0.0674	0.0625	0.0644	0.0643	0.0684
DEU	0.0465	0.0515	0.0492	0.0467	0.0444	0.0484
CHN	0.0393	0.0363	0.0385	0.0383	0.0409	0.0398
JPN	0.0304	0.0291	0.0319	0.0296	0.0273	0.0320
BRA	0.0492	0.0551	0.0479	0.0510	0.0520	0.0480
VNM	0.0592	0.0574	0.0631	0.0585	0.0578	0.0626
GBR	0.0614	0.0667	0.0608	0.0618	0.0597	0.0633
KOR	0.0371	0.0354	0.0378	0.0357	0.0349	0.0382
RUS	0.0383	0.0489	0.0389	0.0374	0.0383	0.0357
NGA	0.0522	0.0568	0.0506	0.0525	0.0514	0.0538
Mean	**0.0479**	0.0505	0.0481	0.0476	0.0471	0.0490
Δ vs D0	-	$+0.0025$	$+0.0002$	-0.0003	-0.0008	$+0.0011$

D1-NoAug shows the largest degradation ($+0.0025$), justifying synthetic augmentation of the under-represented Species and Utilitarianism categories. D4-NoFlip is marginally lower than D0, indicating that group-side randomisation is conservative (it adds noise) rather than harmful; we retain it because it preserves positional-debiasing integrity. The 0.0034 aggregate range is well below the ± 0.004 bootstrap CI, so reported results are not confounded by preprocessing choices.

A17 Trigger Mechanism and Latency

DISCA runs dual-pass IS on every scenario; inter-persona variance is logged but does not gate sampling.

A latency benchmark on one H100 GPU shows similar overhead across proposal budgets: 0.086s/scenario at $K{=}64$, 0.083s at $K{=}128$, and 0.083s at $K{=}256$, compared with 0.023s for vanilla decoding (3.57 to 3.72$\times$ overhead).

Table 34: Inference latency benchmark (Phi-4, H100).

Method	K	sec/scenario	Overhead
DISCA	64	0.086	3.72$\times$
DISCA	128	0.083	3.59$\times$
DISCA	256	0.083	3.57$\times$
Vanilla	n/a	0.023	1.00$\times$

Cost-vs-quality frontier across backbones. Figure 5 plots the per-scenario latency logged during the headline runs against DISCA's macro MIS. These runs used different hardware and quantisation per backbone, so the figure shows orders of magnitude, not a controlled comparison; the controlled overhead is the single-H100 benchmark above. Phi-4 with DISCA (350 ms per scenario, MIS 0.346) is both faster and better aligned than Llama-3.3-70B with DISCA (1414 ms, MIS 0.668). Qwen2.5-7B is faster still (152 ms) at a slightly higher MIS (0.362).

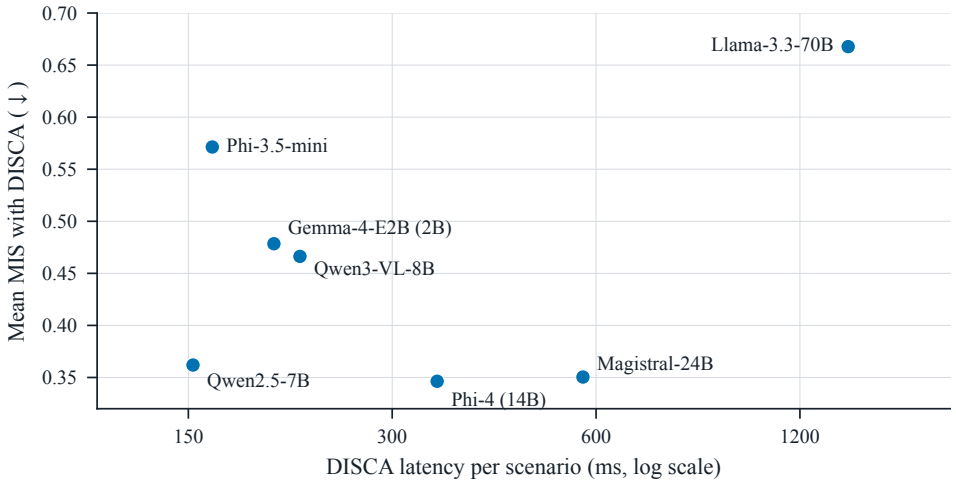


Figure 5: **Cost against quality for the seven headline backbones.** Mean per-scenario DISCA latency logged during the headline runs (log scale) against mean DISCA MIS over the 20 countries. Phi-4 has the lowest MIS and is faster than both Magistral-24B and Llama-3.3-70B; Qwen2.5-7B is faster still at a slightly higher MIS.

A18 Relationship to Persona-Dependent LLM Alignment

Kim et al. [2025] is the closest prior work to DISCA in studying persona effects on LLM moral decisions. Their study and ours share the Moral Machine framework and AMCE-based evaluation, but differ fundamentally in goal, method, and scope. Table 35 summarises the key distinctions.

Table 35: Comparison between Kim et al. [2025] and DISCA.

Dimension	Kim et al. (2025)	DISCA (this work)
Goal	Measure persona sensitivity	Correct cultural misalignment
Personas	7 sociodemographic categories	4 WVS-grounded agents per country
Persona source	Author-designed binary labels	Empirical WVS-7 microdata
Metric	MDD (persona-pair distance)	MIS (ℓ_2 to human AMCE)
Correction	None (diagnostic only)	Consensus, anchor, country prior, gated PT-IS
Models	3 (GPT-4o, GPT-3.5, Llama-2)	7 checkpoints, 5 families, 2B–70B
Countries	Aggregate (Western vs. Eastern)	20 individual countries
Scenarios	9 dimensions, 10k synthetic	6 dimensions, 310 per country
Theory	Partisan sorting (descriptive)	Oracle shrinkage motivating a heuristic (Prop. 1)

Three findings from Kim et al. [2025] directly inform DISCA's design.

Persona sensitivity motivates disagreement-driven steering. Kim et al. show that contrasting personas move LLM moral decisions further apart than the corresponding demographic groups move human decisions, most strongly for political personas. This confirms that persona prompts do shift LLM moral reasoning substantially, validating the premise that within-country persona spread is an informative signal. DISCA converts this sensitivity from a liability (uncontrolled decision shifts) into a feature (a key input to correction reliability, which motivates a shrinkage heuristic; Proposition 1 gives the oracle form).

The partisan sorting phenomenon justifies WVS grounding. Kim et al. find that political orientation produces the largest persona effect in LLMs, far exceeding its effect in human respondents. This "partisan sorting" effect reveals that LLMs over-index on politically salient dimensions when given sociodemographic labels. DISCA avoids this pathology by grounding personas in empirical WVS-7 microdata rather than author-designed binary labels: the ten-dimensional cultural profile (Table 28) spreads influence across religiosity, child-rearing values, social trust, and other axes, so

no single dimension monopolises the correction. The leave-one-WVS-dim-out probe (Table 30) confirms that influence is spread across multiple WVS axes rather than monopolised by one-individual cells can be large in either direction, but the load-bearing dimensions vary by MultiTP attribute, with no single WVS descriptor dominating across the board.

Moral flips validate the reliability gate. Kim et al. document that specific personas can entirely reverse LLM preferences on certain moral dimensions (e.g., social status under a progressive persona in GPT-4o). DISCA's dual-pass reliability gate (Eq. 5) is designed precisely for this regime: when persona-driven corrections are large enough to flip the decision, the two independent IS passes are likely to disagree, and the exponential decay weight r shrinks the correction toward zero. We do not test this mechanism in isolation, so we do not claim that the gate protects against persona-induced flips.

From diagnosis to correction. The fundamental advance of DISCA over the Kim et al. framework is operational: where they diagnose persona sensitivity, we harness it. Their MDD metric measures the distance between two contrasting personas; our MIS measures the distance to the human target and actively reduces it. The variance-aware shrinkage result (Proposition 1) provides the formal bridge: within-panel disagreement (analogous to their MDD) is a key input to the correction's reliability and motivates a shrinkage heuristic (Proposition 1 gives the oracle form).

A19 Extended Limitations

This appendix expands the limitations summarised in §5 with technical caveats not covered in the main text.

WVS-to-trolley linkage. The mapping from WVS value dimensions to trolley moral dimensions operates through both direct thematic overlap and indirect modulation (App. A14.3). The 10-dimensional WVS profile explains $R^2 \in [0.55, 0.69]$ of human AMCE variance, and a leave-one-dimension-out probe identifies load-bearing couplings alongside noisy ones. We avoid claims of the form "WVS dimension X caused AMCE shift Y," and we do not claim that the controller filters out noisy descriptors.

Inter-persona reward assumption. The reward $r_i = \delta_i - \delta_{\text{base}}$ assumes persona shifts approximate human target directions. This is plausible when the base model represents a WEIRD-biased prior, but if a persona's shift is orthogonal to the human target, the signal may mislead. Quantifying whether persona-consensus directions correlate with country-level human AMCE vectors independently of the base model, and when corrections diverge from targets, is open; we do not report that correlation matrix here.

Geographic coverage, preprocessing, and WVS vintage. We report 20 countries out of 100+ MultiTP countries; results depend on the preprocessing pipeline in App. A16. WVS Wave 7 ($\sim$2017 to 2022) is a fixed historical slice that may not reflect current preferences in rapidly evolving societies; persona quality is weakest where coverage is sparse (e.g., ETH and BGD in our panel; Saudi Arabia is excluded entirely and uses a manual fallback persona, App. A14.2). Only two panel languages are low-resource, Amharic (Ethiopia) and Burmese (Myanmar); other low-resource languages such as Khmer or Yoruba remain untested, and one language per country may conflate linguistic and cultural effects.

Model quantisation. 4-bit and Int8 quantisation can introduce logit artefacts that affect importance-sampling stability; we observed this on an early 4-bit Phi-4 run. The Llama-3.3-70B and Qwen3-VL-8B headline runs use 4-bit checkpoints and were not rerun in BF16, so part of their behaviour may reflect quantisation.

Claims this evidence does not support. Table 36 lists stronger claims that a reader might draw from the results, the evidence we do have, and why it falls short.

Table 36: Claims outside the current evidence.

Claim not made	Evidence present	Why the stronger claim is not supported
Demographic worst-group fairness	Synthetic-persona tail statistics and a per-persona utility floor	Synthetic personas are not demographic groups, and no country-by-demographic evaluation is reported.
Cohort-level persona faithfulness	WVS-AMCE associations, WVS feature dropouts, the persona-order audit, and the wrong-country profile test	These show that profile content matters. No matched human response set validates each cohort persona as a faithful portrait, and a country-label permutation placebo does not separate countries.
Safe deployment in healthcare, education, moderation, religion, labour, or family decisions	NormAd social-norm classification	Mentioning a domain is not a deployment evaluation; there is no domain-specific safety, expert, or affected-user study.
Opt-in individual personalisation	Country-level targets only	No experiment compares an individual profile with the country aggregate. Country must not be used to infer personal values.
Improved factual cultural knowledge	BLEnD over 16 countries and 13 languages	The aggregate BLEnD result is negative (SEM-B falls by 4.4 points).
Universal superiority over persona averaging	Same-logit controls, NormAd, GlobalOpinionQA	The same-logit controls cover two backbones and do not give the standard rules the country prior. Given the same prior, plain consensus is at least as good as DISCA on six of seven backbones (Table 3). Probability averaging is within 0.004 of DISCA on NormAd and better on GlobalOpinionQA.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and §1 state three claims: (i) DISCA is a training-free inference-time controller that requires no per-country reward models or extra training data; (ii) it cuts macro MIS by 10 to 24% on six of seven open-weight backbones (3.4% on the 2B model) across 20 countries and 6 moral dimensions; (iii) with DISCA, a 14B backbone (Phi-4) ends closer to the human preferences than a vanilla 70B backbone (Llama-3.3-70B). All three are substantiated in Table 2 and Figure 3, and Table 3 shows which components carry the gain. The comparisons with training-free baselines and with same-logit aggregation rules applied without the country prior, and the NormAd transfer are substantiated in Tables 1 and 4 and Appendices A12.4 and A11; scope and assumptions (binary dilemmas with decision-token logits, WVS-7 vintage) are made explicit in §5 and Appendix A19.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §5 discusses personas as statistical aggregates, country as a coarse proxy for culture, alignment versus legitimacy, the remaining minority risk, prohibited individual uses, the transductive country prior, which a component decomposition (Table 3) shows carries the gain together with the persona panel, and the decision-token requirement; Appendix A10 reports a label-picked temperature control and the finer comparisons between the decomposition arms, and Appendix A11 reports the transfer conditions where DISCA does not beat vanilla or probability averaging. Appendix A19 expands these with: WVS-to-trolley linkage being indirect, the inter-persona reward assumption, geographic coverage (20 of 100+ MultiTP countries; sparse personas for under-sampled countries), WVS Wave 7 vintage, and 4-bit/Int8 quantisation artefacts on IS stability.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might

not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main theoretical result, Proposition 1 (the MSE-optimal scalar shrinkage of the consensus correction is a closed-form function of within-panel variance), is stated in the main text with all assumptions (i.i.d. persona noise, ℓ_2 risk). The full proof is given in Appendix A2; the appendix also states three supporting theorems with proofs. The connection to Eq. 3 appears in §3.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §3 and App. A1 give the full algorithm (Algorithm 1, Eqs. 1 to 8); the full hyperparameter table (Table 21) lists every numerical default; the defaults were checked on a three-backbone, five-country prototyping panel and then frozen (Appendix A13). All backbones are publicly available open-weight checkpoints. Preprocessing (cap=80, oversample=36), seeds (seed 42 for every headline run; seeds 42, 101 and 2026 for the two Qwen2.5 backbones), and per-country preprocessing details are documented in Appendices A16, A9.1. The code that produced the reported numbers, the tagged release `neurips2026-camera-ready`, will be made public with the paper; it is a self-contained reference implementation with a single-command sweep entry point.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code that produced the reported numbers (the tagged release `neurips2026-camera-ready`, MIT-licensed) will be made public with the paper. The package is a self-contained slice of the research codebase: a single CLI entry point (`run.py`) drives the per-model 20-country sweep, with backend selection (`vllm` / `hf_native` / `unsloth`) and explicit flags for the three external data paths (`--multitp-data`, `--wvs-data`, `--human-amce`). The README documents install (Python 3.10+, CUDA 12.x), HF token wiring for gated models, exact commands, and the output schema (per-country baseline / DISCA CSVs and a comparison file). MultiTP, WVS-7 Wave, and the human AMCE table are publicly available datasets cited in §4; access instructions are reproduced in the README.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: There is no training (the method is inference-time on frozen weights); we therefore detail all decoding and controller settings. §4 specifies the dataset (MultiTP, 20 countries), preprocessing (deduplication, category cap=80, oversample to $\geq$36 per dimension), the seven main backbones with sizes, and the metric definitions (MIS, Pearson r, JSD, win rate). All hyperparameters are tabulated in Table 21; Appendix A13 states how they were chosen (a three-backbone, five-country prototyping panel, then frozen). Sensitivity sweeps for the temperature family (Appendix A13), category cap (Appendix A16), and the controller knobs (s, λ_{coop}, σ, T_{cat}) appear in Appendix A13.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The headline table (Table 2) reports one run per backbone (seed 42) together with per-country win counts, and every per-country value is given in the appendix. Variability is reported where we have repeated runs: the same-logit controls, the GlobalOpinionQA controller and the persona-tail statistic use three seeds on two Qwen2.5 backbones (Appendices A9.1 and A11), and a bootstrap confidence interval on JSD (± 0.004, Appendix A13) gives the noise floor of the sensitivity sweeps. We do not report error bars for the seven-backbone headline sweep.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix A17 (Table 34) reports per-scenario wall-clock on a single H100 GPU: 0.083 to 0.086 s/scenario for DISCA at $K \in \{64, 128, 256\}$ vs. 0.023 s/scenario for vanilla decoding (3.57 to 3.72$\times$ overhead). Backbone-level cost is given in Figure 5: 350 ms/scenario for Phi-4 vs. 1414 ms for Llama-3.3-70B with DISCA, measured in a different setting from the single-H100 benchmark, so only the comparison between backbones is meaningful. Llama-3.3-70B and Qwen3-VL-8B run from 4-bit quantised checkpoints and Qwen2.5-7B runs in BF16; the precision of the other four runs follows the loader default and was not logged separately. The released code's CLI also runs end-to-end on a node with two T4 or P100 GPUs for sub-14B models. Exploratory experiments on a small 3-backbone $\times$ 5-country prototyping panel consumed additional compute beyond the reported runs. Experiments were run on the free GPU quota provided by Kaggle and on a single H100 GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics. The work uses only public, ethically-collected datasets (MultiTP [Jin et al., 2025], World Values Survey Wave 7, Moral Machine human AMCEs [Awad et al., 2018]); it conducts no new human-subjects research and introduces no scraped data. Negative-impact considerations (risk of encoding harmful majorities) are explicitly discussed in §5. The paper also describes an optional per-persona utility floor (Appendix A13, Table 27), which leaves macro MIS unchanged; it is a heuristic, and we do not show that it protects the worst-off persona.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: §5 discusses both directions. Positive: a training-free controller, with a measured overhead of 3.6 to 3.7 times vanilla inference, lets practitioners respect cross-national preference heterogeneity at deployment time without finetuning per country, lowering the

barrier to culturally pluralistic alignment. Negative: aligning to majority survey statistics could marginalise minority views or entrench stereotyped views of a population, as LLMs already do when they estimate other countries' moral values [Zewail et al., 2026], and alignment to a survey statistic is not equivalent to legitimacy. We describe an optional per-persona utility floor that limits how far any single persona's post-correction utility can drop from vanilla (Appendix A13, Table 27); it leaves macro MIS unchanged, it is a heuristic, and we do not show that it protects the worst-off persona. We are explicit that DISCA is a research artefact, not a deployment-ready normative system.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We do not release any pretrained models, generative models, or scraped datasets. The DISCA code release is an inference-time controller layered over publicly available open-weight LLMs whose existing safeguards remain in place. The method itself produces no novel generative capability beyond that of the underlying models; the optional per-persona utility floor (Appendix A13, Table 27) is a heuristic, not a demonstrated safeguard for the worst-off persona.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification:

All third-party assets are cited with version information. Datasets: MultiTP [Jin et al., 2025] (MIT License, `https://github.com/causalNLP/MultiTP`); the Moral Machine human data [Awad et al., 2018] (OSF project `https://osf.io/3hvt2`, released by its authors for research reuse, with no formal licence); World Values Survey Wave 7 (inverted CSV `v6_0`), used under the WVS data-use conditions (free registration, non-commercial use, no redistribution); NormAd [Rao et al., 2025] (CC BY 4.0); GlobalOpinionQA [Durmus et al., 2024] (CC BY-NC-SA 4.0); and BLEnD [Myung et al., 2024] (CC BY-SA 4.0). Models: Phi-4 and Phi-3.5-mini (MIT), Llama-3.3-70B (Llama 3.3 Community License), Qwen-2.5 / Qwen3-VL (Apache-2.0 / Qwen license), Gemma-4-E2B (Apache-2.0) and the other Gemma models of App. A7 (Gemma terms of use), Magistral-Small-2509 (Apache-2.0). Code dependencies (Transformers, vLLM, Unsloth, bitsandbytes) are listed with version pins in the released `requirements.txt`; we use them under their stated licenses (mostly Apache-2.0 / MIT). The `compute_ACME` reference implementation we follow is credited to Jin et al. [2025] in Appendix A15.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The DISCA reference implementation is a new code asset under the MIT License and will be made public with the paper. The package ships with: a top-level README documenting installation, backend selection (vLLM / HF native / Unsloth), the three external data flags, environment variables, the quick-start command, and the output schema; a flat `disca/` module with one-purpose files (controller, dual-pass core, runner, AMCE/metrics, persona generation, multilingual scenarios, model loader); pinned dependencies in `requirements.txt`; and a single CLI entry point (`run.py --help`). The released code is the tagged version `neurips2026-camera-ready`, which produced the reported numbers with the defaults of Table 21.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper conducts no new crowdsourcing or human-subjects research. Human reference data is reused from the existing public Moral Machine experiment [Awad et al., 2018] as packaged by MultiTP [Jin et al., 2025], and from the World Values Survey Wave 7 public microdata; we make no new contact with participants and recruit no annotators.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new human-subjects research is conducted. The work is a secondary analysis of pre-existing public datasets (Moral Machine human AMCEs, MultiTP scenarios, WVS-7 microdata), each of which carries its own ethics review at the original collection point; no further IRB approval applies.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

Justification: LLMs are central to the work as its subject: the frozen open-weight LLMs (Phi-4, Llama-3.3-70B, Magistral-24B, Qwen3-VL-8B, Qwen2.5-7B, Phi-3.5-mini, Gemma-4-E2B) are what DISCA steers; their identities, sizes, and decoding settings are documented in §4 and Appendix A7. LLMs were not used to generate the paper's claims, results, or proofs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.