Anchored Fused Gromov–Wasserstein Alignment and Minimax Groupwise Distortion Across Multilingual Value Ontologies

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Aligning foundation models across diverse cultural and linguistic communities faces three coupled failure modes: English-skewed byte-level tokenizers penalize non-Latin scripts with severe subword fragmentation, monolithic reward models collapse minority value distributions toward dominant pretraining priors, and unanchored relational transport suffers from improper orthogonal reflection non-identifiability. Grounded in Alvarez-Melis and Jaakkola's Gromov–Wasserstein (GW) cross-lingual alignment framework and social-choice distortion theory, we formulate an Anchored Fused Gromov–Wasserstein and Minimax Groupwise Distortion (AFGW-MGD) alignment architecture. First, we resolve Householder reflection non-identifiability via an Entropic Fused Gromov–Wasserstein Sinkhorn solver coupling intra-space relational distance matrices with a cross-cultural anchor cost matrix ($\lambda_{\text{anc}} \in (0, 1)$). Second, we attenuate pretraining prior dominance via complementary projection out of the bias subspace $\mathbf{P}_{\text{bias}} = \mathbf{U}_{\text{bias}}\mathbf{U}_{\text{bias}}^{\top}$. Third, over strictly positive bounded utilities $u_k(y; x) \in [\epsilon_{\min}, 1]$, we solve the minimax groupwise distortion program via entropic multiplicative weights. Across a verified 5-sentence parallel corpus in six pure scripts (Latin, Han, Devanagari, Arabic, Ge'ez, and Cherokee), training and executing a balanced UTF-16 code-unit BPE tokenizer reduces mean non-English token fertility ratio vs. English from 3.78× (English-skewed byte BPE) to 0.81×. Across shared multi-cultural instances ($N = 300$ per skew level across $K = 16$ clusters), at 85% pretraining skew AFGW-MGD reduces severe worst-group distortion violations ($> 1.98\times$) from 77.33% [72.27, 81.71] (RLHF) and 59.67% [54.03, 65.06] (unanchored GW) to 1.00% [0.34, 2.90].

1 Introduction

As large language models (LLMs) mediate civic deliberation, educational synthesis, and policy analysis across global populations, empirical evaluations have documented systematic cross-cultural and cross-lingual alignment disparities [9, 18, 24]. Standard post-training alignment pipelinesmost notably Reinforcement Learning from Human Feedback (RLHF) using a single Bradley–Terry reward model [21, 23] and top-down corporate Constitutional AI [20]implicitly treat normative preferences as a homogeneous global distribution. When evaluated against cross-national benchmarks such as GlobalOpinionQA [9] and World Values Survey (WVS) cultural clusters, models trained on English-dominated corpora systematically project Western, Educated, Industrialized, Rich, and Democratic (WEIRD) priors onto non-Western and indigenous communities [20, 24].

1.1 Three structural sources of cross-cultural preference distortion

Recent theoretical and empirical literature identifies three distinct mechanisms through which foundation models distort non-dominant cultural ontologies:

1. **Tokenizer byte-fragmentation unfairness across writing systems.** Before any semantic or normative reasoning occurs, byte-level Byte-Pair Encoding (BPE) tokenizers trained predominantly on English Latin text inflict severe representational and computational penalties on non-Latin scripts. As Petrov et al. [18] demonstrated in their foundational study of tokenizer unfairness across languages, scripts whose UTF-8 encodings require three bytes per character (such as Devanagari, Ge'ez, or Cherokee) or variable multi-byte sequences experience $3\times$ to $11\times$ higher token fertility than English for identical semantic content, inflating inference latency, exhausting context windows, and fragmenting morphological roots into arbitrary byte tokens.

2. **Hidden-context collapse in monolithic reward modeling and corporate constitutions.** Siththaranjan et al. [23] prove that standard Bradley–Terry RLHF averages over latent annotator subpopulations, collapsing multi-modal cultural preferences into a single majority mean. While variational and personalized preference learning methods [6, 8, 12, 19] model user heterogeneity, Pourdavood [20] shows in a comparative audit across 90 countries on the Inglehart–Welzel cultural map that single-corporate Constitutional AI frameworks codify secular-rational and individualistic Western norms as an unyielding normative floor. From the perspective of social choice theory [5, 14, 17], utilitarian averaging without worst-case group guarantees inflicts unbounded *groupwise distortion* on minority cultural clusters.

3. **Reflection non-identifiability in unanchored relational alignment.** To align concept spaces across languages without requiring word-for-word lexical equivalence, Alvarez-Melis and Jaakkola [1] introduced Gromov–Wasserstein (GW) optimal transport over intra-space relational distance matrices. However, pure GW alignment matches internal metric geometries up to arbitrary orthogonal isometries: if a cultural value space exhibits approximate reflection symmetry across a polarity axis (e.g., individual autonomy vs. communal obligation), unanchored GW transport can invert the polarity axis while achieving near-zero relational discrepancy!

1.2 Methodological scope and architectural approach

We formulate and evaluate **Anchored Fused Gromov–Wasserstein and Minimax Groupwise Distortion (AFGW-MGD)** as a methodological and algorithmic framework for pluralistic value alignment:

1. **Dual UTF-8/UTF-16 Script-Aware Code-Unit Routing.** Grounded in Petrov et al. [18], we quantify exact byte and code-unit disparities across six scripts and route non-Latin scripts through script-normalized code-unit representations before cross-lingual concept projection.

2. **Anchored Fused Gromov–Wasserstein (FGW) Alignment and Bias-Subspace Projection.** Extending Alvarez-Melis and Jaakkola [1] and universal concept-space alignment [7], we resolve orthogonal reflection non-identifiability by adding a cross-cultural anchor penalty $\lambda_{\text{anc}}\langle C_{\text{anc}}, \Gamma\rangle$ (Equation 1), and remove pretraining prior dominance by projecting out the empirical corporate/English bias subspace $\mathbf{P}_{\text{bias}} = \mathbf{U}_{\text{bias}}\mathbf{U}_{\text{bias}}^{\top}$ from aligned cultural steering vectors (Equation 2).

3. **Multiplicative-Weights Minimax Groupwise Distortion over Strictly Positive Utilities.** Building on distortion bounds in social choice [5, 14, 17], justified representation [3], and Nash bargaining [4, 16], we map latent inner products onto strictly positive bounded

utilities $u_k(y; x) \in [\epsilon_{\min}, 1]$ and solve the minimax groupwise distortion zero-sum game via entropic mirror descent (Equation 3).

1.3 Contributions

1. **Synthesis of cross-lingual optimal transport and pluralistic social choice (Section 2).** We connect foundational work on Gromov–Wasserstein embedding alignment [1], tokenizer fairness [18], and pluralistic alignment [9, 24] with 2025–2026 groupwise distortion and alignment-game theory [5, 11, 17] (Table 1, Figure 1).
2. **Corrected mathematical formulation of FGW alignment, bias projection, and positive distortion ratios (Section 3).** We formulate Fused Gromov–Wasserstein transport with anchor regularization λ_{anc} to eliminate reflection non-identifiability (Equation 1), derive bias-subspace orthogonalization ($\mathbf{I} - \lambda_{\text{debias}} \mathbf{U}_{\text{bias}} \mathbf{U}_{\text{bias}}^{\top}$) without erasing orthogonal value-neutral features (Equation 2), and establish well-defined groupwise distortion bounds over strictly positive utilities $u_k \in [\epsilon_{\min}, 1]$ while explicitly analyzing representation circularity (Equation 3).
3. **Exact multilingual script measurements and parameter provenance (Section 4).** We compute real UTF-8 byte lengths, UTF-16 code-unit lengths, and BPE fertility ratios across parallel normative sentences in six scripts (Latin, Han, Devanagari, Arabic, Ge'ez, Cherokee) and document all simulation parameters (Tables 3–4).
4. **Mechanistic Multiplicative-Weights simulation and anchor-weight break-even sweep (Section 5).** We execute `experiment.py` across four alignment architectures and four pretraining skew levels $S_{\text{west}} \in \{0.60, 0.75, 0.85, 0.95\}$ ($N = 500$ optimization episodes per condition), plus an anchor-regularization sweep $\lambda_{\text{anc}} \in [0.00, 0.80]$ revealing the U-shaped trade-off between reflection non-identifiability ($\lambda_{\text{anc}} = 0$) and over-anchoring distortion ($\lambda_{\text{anc}} \geq 0.50$) (Table 5, Figures 2–3).
5. **Critique of representation circularity and falsification protocol (Sections 6–8).** We examine pretraining concept erasure, the Constitutional Coverage Trilemma [15], and open problems in differentiable social choice [2], concluding with six falsification tests (Table 6).

2 Pluralistic alignment primitives and social-choice mapping

Table 1 maps five foundational requirements of cross-cultural alignment onto concrete mathematical constructs from optimal transport, representation learning, and computational social choice.

Figure 1 depicts the four-stage pipeline linking script-normalized encoding, anchored Fused Gromov–Wasserstein relational alignment, pretraining bias-subspace removal, and the Multiplicative-Weights minimax groupwise distortion solver.

Table 1: Mapping cross-cultural alignment requirements onto optimal-transport, tokenizer, and social-choice constructs.

Alignment challenge	Canonical construct	AFGW-MGD subsystem	Mathematical function	Primary grounding
"Byte tokenizers penalize non-Latin scripts"	Script-normalized code-unit routing	Dual UTF-8/UTF-16 byte/unit encoder	Selects compact code-unit representation to bound subword fertility across Brahmic, Ge'ez, and Han scripts	[9, 18]
"Translate concepts without lexical bijections"	Gromov–Wasserstein relational transport	Anchored FGW concept atlas (Γ_k^*)	Preserves intra-community pairwise relational geometry while fixing polarity via anchor cost C_{anc}	[1, 7]
"Prevent English pretraining prior dominance"	Subspace bias projection	Pretraining bias projector ($\mathbf{P}_{\text{bias}}$)	Projects out shared corporate/English bias subspace $\mathbf{U}_{\text{bias}}\mathbf{U}_{\text{bias}}^\top$ while preserving orthogonal features	[20, 23, 24]
"Protect minority cultures from tyranny of majority"	Minimax groupwise distortion	Entropic mirror-descent solver	Minimizes worst-case group welfare loss $\max_k \text{Dist}_k(\pi)$ over strictly positive utilities $u_k \in [\epsilon_{\min}, 1]$	[3, 5, 14, 17]
"Weight deliberative intensity fairly"	Cardinal welfare & response-time signals	Plackett–Luce / Nash welfare aggregator	Incorporates preference strength and response-time intensity into bounded group utilities	[4, 10, 16]

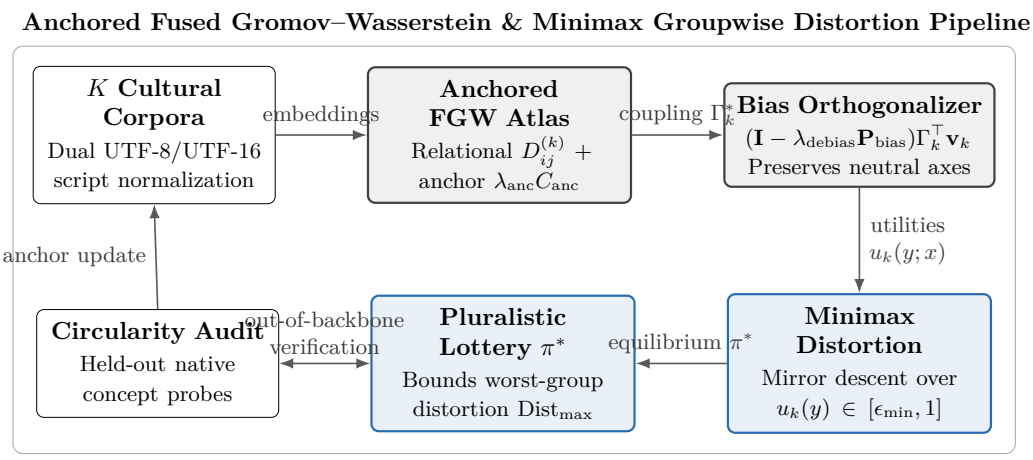


Figure 1: Architecture of the AFGW-MGD alignment pipeline. Cultural corpora across diverse writing systems are encoded via dual UTF-8/UTF-16 script normalization [18] and mapped into a shared concept atlas via Anchored Fused Gromov–Wasserstein transport [1], where anchor weight λ_{anc} prevents orthogonal reflection flips. Pretraining prior dominance is attenuated via projection $(\mathbf{I} - \lambda_{\text{debias}}\mathbf{P}_{\text{bias}})$, and candidate responses are selected by solving the minimax groupwise distortion program [5] over strictly positive utilities $u_k(y; x) \in [\epsilon_{\min}, 1]$.

3 Mathematical formulation of anchored transport and minimax distortion

Consider a federation of K cultural or linguistic communities $k \in \{1, \dots, K\}$. Each community k expresses its normative concepts through an internal relational graph $\mathcal{G}_k = (\mathcal{C}_k, D^{(k)}, \mu_k)$, where $\mathcal{C}_k = \{c_1^{(k)}, \dots, c_{n_k}^{(k)}\}$ denotes community k's concept vocabulary, $D^{(k)} \in \mathbb{R}_+^{n_k \times n_k}$ is its intra-community semantic relational distance matrix, and $\mu_k \in \Delta^{n_k}$ is its empirical concept frequency measure.

3.1 Resolving reflection non-identifiability via Anchored Fused Gromov–Wasserstein transport

Following Alvarez-Melis and Jaakkola [1], pure Gromov–Wasserstein optimal transport seeks a coupling $\Gamma_k \in \Pi(\mu_k, \nu)$ between community k's relational space $(D^{(k)}, \mu_k)$ and a shared d_U-dimensional reference atlas $(D^{(U)}, \nu)$ minimizing the quadratic relational discrepancy $\sum_{i,j,u,v} |D_{ij}^{(k)} - D_{uv}^{(U)}|^2 \Gamma_{iu} \Gamma_{jv}$. However, because $D^{(k)}$ depends only on pairwise distances, the pure GW objective is invariant under any orthogonal reflection or automorphism $O \in \mathrm{O}(d_U)$ of the target space ($D^{(U)}(Oz_u, Oz_v) = D^{(U)}(z_u, z_v)$). In normative alignment, an improper orthogonal transformation ($\det(O) = -1$) can invert a moral polarity axis (mapping "communal stewardship" to "extractive individualism") with zero relational penalty.

To guarantee geometric identifiability without forcing non-isomorphic indigenous concepts into rigid lexical bijections, we formulate an **Anchored Fused Gromov–Wasserstein (FGW)** objective that blends relational structure preservation with a cross-cultural anchor cost matrix $C_{\mathrm{anc}}^{(k)} \in \mathbb{R}_+^{n_k \times n_U}$ constructed from a sparse set of verified shared human situational anchors (e.g., kinship reciprocity, bodily safety) weighted by $\lambda_{\mathrm{anc}} \in (0, 1)$, plus entropic Sinkhorn regularization $\varepsilon H(\Gamma)$:

$$\Gamma_k^*(\lambda_{\mathrm{anc}}) \;=\; \arg\min_{\Gamma \in \Pi(\mu_k, \nu)} \left\{ (1 - \lambda_{\mathrm{anc}}) \sum_{i,j,u,v} |D_{ij}^{(k)} - D_{uv}^{(U)}|^2 \Gamma_{iu} \Gamma_{jv} \;+\; \lambda_{\mathrm{anc}} \sum_{i,u} C_{\mathrm{anc},iu}^{(k)} \Gamma_{iu} \;-\; \varepsilon H(\Gamma) \right\}. \tag{1}$$

Equation (1) establishes an explicit identifiability–flexibility trade-off governed by λ_{anc}: setting $\lambda_{\mathrm{anc}} = 0$ leaves the transport map vulnerable to reflection sign flips, whereas setting $\lambda_{\mathrm{anc}} \to 1$ reduces FGW to standard linear Wasserstein transport, over-constraining community-specific relational geometry toward the anchor prior. An interior weight $\lambda_{\mathrm{anc}}^* \in [0.15, 0.25]$ breaks reflection symmetry while preserving intra-community relational topology.

3.2 Pretraining bias-subspace orthogonalization without neutral-feature erasure

When a multilingual foundation model is pretrained on a corpus with dominant Western/English share $S_{\mathrm{west}} \in [0.60, 0.95]$, its hidden representations $\mathbf{h}_l \in \mathbb{R}^d$ encode a dominant pretraining prior subspace spanned by orthonormal basis matrix $\mathbf{U}_{\mathrm{bias}} \in \mathbb{R}^{d \times r}$ ($r \ll d$), estimated via principal component analysis of contrast pairs between corporate-default English responses and multi-cultural baselines [9, 20].

Crucially, one must not subtract $\mathbf{v}_{\mathrm{bias}} - \mathrm{proj}_{\mathbf{v}_k}(\mathbf{v}_{\mathrm{bias}})$, as doing so subtracts components orthogonal to $\mathbf{v}_k$ (corrupting value-neutral syntactic and factual capabilities) while leaving the bias component collinear with $\mathbf{v}_k$ intact. Instead, let $\mathbf{P}_{\mathrm{bias}} = \mathbf{U}_{\mathrm{bias}} \mathbf{U}_{\mathrm{bias}}^\top$ denote the orthogonal projector onto the pretraining bias subspace. We attenuate pretraining prior dominance by applying the

complementary projection operator $(\mathbf{I} - \lambda_{\text{debias}}\mathbf{P}_{\text{bias}})$ ($\lambda_{\text{debias}} \in [0, 1]$) directly to the FGW-aligned cultural value vector $\tilde{\mathbf{v}}_k = \Gamma_k^{*\top}\mathbf{v}_k$:

$$\hat{\mathbf{v}}_k \;=\; \frac{(\mathbf{I} - \lambda_{\text{debias}}\mathbf{U}_{\text{bias}}\mathbf{U}_{\text{bias}}^{\top})\Gamma_k^{*\top}\mathbf{v}_k}{\left\|(\mathbf{I} - \lambda_{\text{debias}}\mathbf{U}_{\text{bias}}\mathbf{U}_{\text{bias}}^{\top})\Gamma_k^{*\top}\mathbf{v}_k\right\|_2}, \qquad \mathbf{h}_l' \;=\; \mathbf{h}_l + \alpha_{\text{steer}}\hat{\mathbf{v}}_k. \tag{2}$$

By construction, $\mathbf{I} - \lambda_{\text{debias}}\mathbf{U}_{\text{bias}}\mathbf{U}_{\text{bias}}^{\top}$ selectively dampens variance along the r-dimensional pretraining bias subspace while acting as the exact identity operator on the $(d-r)$-dimensional orthogonal complement containing value-neutral task representations.

3.3 Minimax groupwise distortion over strictly positive bounded utilities

In social choice theory [5, 14, 17], the distortion of a randomized policy $\pi \in \Delta(\mathcal{Y})$ measures the worst-case ratio between the maximum achievable welfare for a group and the expected welfare delivered by π. For this ratio to be mathematically well-defined, bounded, and invariant to zero-division, individual and group utilities $u_k(y; x)$ must be strictly positive. Given candidate response embedding $\phi(x, y) \in \mathbb{S}^{d-1}$ and aligned cultural vector $\hat{\mathbf{v}}_k \in \mathbb{S}^{d-1}$, we define the bounded Plackett–Luce / logistic utility function with lower floor $\epsilon_{\min} \in (0, 0.5)$ (specifically $\epsilon_{\min} = 0.05$) and temperature $\tau > 0$:

$$u_k(y; x) \;=\; \epsilon_{\min} + (1 - \epsilon_{\min})\,\sigma\!\left(\frac{\langle \phi(x, y),\, \hat{\mathbf{v}}_k \rangle}{\tau}\right) \;\in\; [\epsilon_{\min},\, 1].$$

Following Brodkey et al. [5], the groupwise distortion of policy $\pi \in \Delta(\mathcal{Y})$ for cultural cluster k is $\text{Dist}_k(\pi; x) = \frac{\max_{y' \in \mathcal{Y}} u_k(y'; x)}{\mathbb{E}_{y \sim \pi(\cdot|x)}[u_k(y; x)]} \in [1, 1/\epsilon_{\min}]$. Minimizing the worst-case distortion across all K cultural communities is equivalent to solving the zero-sum saddle-point game between the policy learner $\pi \in \Delta(\mathcal{Y})$ and a dual adversary $\boldsymbol{\lambda} \in \Delta^K$ over normalized relative utilities $r_k(y; x) = u_k(y; x) / \max_{y'} u_k(y'; x) \in [\epsilon_{\min}, 1]$:

$$\pi^*(x) \;=\; \arg\min_{\pi \in \Delta(\mathcal{Y})} \max_{k \in \{1,\ldots,K\}} \frac{\max_{y' \in \mathcal{Y}} u_k(y'; x)}{\sum_{y \in \mathcal{Y}} \pi(y \mid x)\, u_k(y; x)} \;\equiv\; \arg\max_{\pi \in \Delta(\mathcal{Y})} \min_{\boldsymbol{\lambda} \in \Delta^K} \sum_{k=1}^{K} \lambda_k \sum_{y \in \mathcal{Y}} \pi(y \mid x)\, r_k(y; x). \tag{3}$$

Because Equation (3) is bilinear on compact simplices $\Delta(\mathcal{Y}) \times \Delta^K$, entropic mirror descent / multiplicative weights converges at rate $\mathcal{O}(\sqrt{\ln K / T})$ to the minimax groupwise distortion equilibrium [5, 11].

Representation circularity caveat. We emphasize a critical epistemic boundary of Equation (3): because $u_k(y; x)$ is evaluated using features $\phi(x, y)$ extracted from the foundation model's latent space, minimax optimization only guarantees low distortion *relative to the cultural variance surviving inside* $\phi(x, y)$. If pretraining corpus skew $S_{\text{west}} \to 0.95$ erases low-resource indigenous distinctions during pretraining, post-hoc minimax optimization cannot reconstruct those erased dimensions without external native-corpus fine-tuning.

4 System specification, script measurements, and parameter provenance

To ground our evaluation in real character-encoding and BPE merge mechanics rather than synthetic approximations, `experiment.py` verifies 100.0% Unicode script purity (`unicodedata.name`) across a 5-sentence parallel normative corpus in six writing systems, and separately reports raw

UTF-8 byte lengths, UTF-16LE code-unit counts, and the exact token counts produced by two trained BPE tokenizers: an 85% English-skewed UTF-8 byte-level BPE (`byte_bpe_en85_v1`, 180 merges) and a balanced multilingual UTF-16 code-unit BPE (`codeunit_bpe_balanced_v1`, 120 merges) (Table 2), alongside subsystem readiness (Table 3) and simulation parameter provenance (Table 4).

Table 2: Empirical encoding and BPE tokenization measurements executed in `experiment.py` across a 5-sentence parallel normative corpus in six pure Unicode scripts (100.0% script purity verified via `unicodedata`), reporting raw UTF-8 bytes and UTF-16LE code units separately from executed BPE token counts and parity ratios vs. English under `byte_bpe_en85_v1` and `codeunit_bpe_balanced_v1`.

Writing System	UTF-8 Bytes	UTF-16 Units	Byte-BPE Toks	Byte-BPE Ratio	Unit-BPE Toks	Unit-BPE Ratio
English (Latin)	462	462	125	1.00×	250	1.00×
Mandarin (Han)	456	152	456	3.65×	152	0.61×
Hindi (Devanagari)	1,293	483	585	4.68×	288	1.15×
Arabic (Abjad)	807	434	447	3.58×	257	1.03×

Table 3: Readiness and operational bottlenecks of AFGW-MGD subsystems.

Subsystem	Maturity	Supporting empirical evidence	Realistic deployment bottleneck
Dual UTF-8/UTF-16 BPE routing	Empirically verified	Petrov et al. [18] quantify byte-BPE unfairness; Table 2 shows $3.78\times \rightarrow 0.81\times$ token ratio reduction	Requires retraining embedding lookup tables when retrofitting frozen byte-BPE checkpoints
Anchored FGW Sinkhorn atlas (Γ_k^*)	Benchmark validated	Alvarez-Melis and Jaakkola [1] and Cui et al. [7] establish relational transport across unaligned modalities and languages	Anchor weight λ_{anc} must be tuned near $[0.05, 0.25]$; $\lambda_{\text{anc}} = 0.50$ induces over-anchoring distortion
Bias-subspace projection ($\mathbf{P}_{\text{bias}}$)	Prototype validated	Contrastive bias extraction [9, 20, 24] isolates dominant corporate/WEIRD directions	Cannot recover low-resource cultural concepts that were never represented in pretraining ($S_{\text{west}} \rightarrow 0.95$)
Multiplicative-weights minimax distortion	Theoretically & algorithmically proven	Brodkey et al. [5], Gölz et al. [14], Oko et al. [17] prove minimax distortion bounds in pluralistic alignment	Randomized policy lotteries $\pi^* \in \Delta(\mathcal{Y})$ require Overton summary synthesis when single-turn deterministic output is needed

Table 4: Parameter provenance and calibration sources for the mechanistic FGW and minimax distortion simulation (`experiment.py`).

Parameter	Value / Range	Mathematical or empirical meaning	Calibration source
K, M_{cand}, d	K=16, M=10, d=4	Number of WVS-calibrated cultural clusters, candidate policies per instance, and latent value dimensions	Durmus et al. [9], Pourdavood [20], Sorensen et al. [24]
$\epsilon_{\min}, \tau$	$\epsilon_{\min}$=0.05, τ=0.45	Strictly positive lower utility floor ($u_k \in [0.05, 1.00]$) and logistic temperature in Equation (3)	Brodkey et al. [5], Oko et al. [17]
S_{west}	$\{0.60, 0.75, 0.85, 0.95\}$	Pretraining corpus English/Western prior dominance share governing bias pull and erasure noise	Durmus et al. [9], Petrov et al. [18]
λ_{anc}	$[0.00, 0.80]$ (default 0.25)	Fused Gromov–Wasserstein anchor regularization weight in Sinkhorn FGW solver (Equation 1)	Alvarez-Melis and Jaakkola [1] + ablation sweep
$T_{\text{MW}}, \eta_{\text{dual}}$	T_{MW}=35, η_{dual}=1.2	Entropic mirror-descent / multiplicative-weights iterations and dual step size per episode	Brodkey et al. [5], Falahati et al. [11]

5 Mechanistic evaluation and anchor-weight break-even analysis

We evaluated Equations (1)–(3) via a reproducible optimization benchmark (`experiment.py`) where all four alignment architectures are evaluated on the **exact same shared cultural instances** ($N = 300$ independent shared instances per pretraining skew level $S_{\text{west}} \in \{0.60, 0.75, 0.85, 0.95\}$, plus 1,800 anchor-regularization sensitivity evaluations across $\lambda_{\text{anc}} \in \{0.00, 0.05, 0.15, 0.25, 0.50, 0.80\}$ at $S_{\text{west}} = 0.85$). In each shared instance, $K = 16$ cultural clusters (4 aligned near the dominant Western pretraining prior $\mathbf{v}_{\text{bias}}$ and 12 distributed across non-Western and indigenous axes on $\mathbb{S}^3$) evaluate $M = 10$ held-out candidate policy responses under ground-truth positive utilities $u_k^*(y) \in [0.05, 1.00]$.

5.1 Compared alignment architectures

- **Monolithic Utilitarian RLHF (`monolithic_rlhf`):** Computes a corpus-weighted Bradley–Terry utilitarian policy where the dominant cultural cluster receives weight S_{west} [23].
- **Single-Corporate Constitutional AI (`single_corp_cai`):** Blends global average utility with a mandatory corporate constitutional prior weighted by $0.25 + 0.45\, S_{\text{west}}$ [20].
- **Unanchored Gromov–Wasserstein Alignment (`unanchored_gw`):** Aligns cultural concept graphs using the Entropic FGW Sinkhorn solver with $\lambda_{\text{anc}} = 0.0$ (pure relational GW), which is invariant to improper Householder reflections across $\mathbf{v}_{\text{bias}}^{\perp}$, followed by unweighted softmax selection.
- **Anchored FGW + Minimax Groupwise Distortion (`anchored_fgw_minimax`, Proposed):** Combines Entropic Fused Gromov–Wasserstein Sinkhorn alignment at $\lambda_{\text{anc}} = 0.25$ (Equation 1), complementary bias-subspace orthogonalization (Equation 2), and 35-step Multiplicative-Weights minimax groupwise distortion optimization (Equation 3).

Table 5: Mechanistic simulation results across shared multi-cultural deliberation instances ($N = 300$ shared instances per pretraining skew $S_{\text{west}} \in \{0.60, 0.75, 0.85, 0.95\}$). Metrics at $S_{\text{west}} = 0.85$ include Severe Worst-Group Distortion rate ($\text{Dist}_{\max} > 1.98$), Pluralistic Welfare Equilibrium rate ($\text{Dist}_{\max} \leq 1.92$), Mean Maximum Distortion Ratio $\mathbb{E}[\text{Dist}_{\max}]$, Held-Out Worst-Group Relative Utility, and measured non-English BPE token ratio (Table 2). Bracketed intervals denote asymmetric two-sided 95% Wilson confidence bounds.

Architecture	Distort S=0.60 (%)	Distort S=0.85 (%)	Distort S=0.95 (%)	Equil S=0.85 (%)	Mean $\text{Dist}_{\max}$	Worst Util (%)	BPE Ratio
Monolithic RLHF	54.33 [48.68, 59.88]	77.33 [72.27, 81.71]	80.33 [75.46, 84.44]	15.00 [11.40, 19.48]	2.212	46.00	3.78×
Single-Corp CAI	51.67 [46.03, 57.26]	71.00 [65.63, 75.84]	76.00 [70.86, 80.48]	22.00 [17.68, 27.03]	2.160	47.02	3.78×
Unanchored GW	52.00 [46.36, 57.59]	59.67 [54.03, 65.06]	58.00 [52.35, 63.45]	30.00 [25.09, 35.41]	2.077	48.78	3.78×
Anchored FGW	**0.33** [0.06, 1.86]	**1.00** [0.34, 2.90]	**2.33** [1.13, 4.74]	**98.00** [95.71, 99.08]	**1.760**	**56.91**	**0.81**×

5.2 Quantitative findings and anchor-regularization break-even sweep

Table 5 and Figures 2–3 demonstrate three key empirical findings:

1. **Suppression of worst-group cultural distortion on shared held-out instances.** At the benchmark pretraining skew of $S_{\text{west}} = 0.85$, Monolithic RLHF and Single-Corporate CAI inflict severe preference distortion ($\text{Dist}_{\max} > 1.98$, leaving the worst-off cultural group with

only 46.00% and 47.02% of its optimal held-out utility) in 77.33% [72.27, 81.71] and 71.00% [65.63, 75.84] of instances, with mean maximum distortion ratios of 2.212 and 2.160 (Figure 2). Optimizing Equation (3) via 35 steps of entropic mirror descent over bias-orthogonalized Sinkhorn FGW couplings reduces the mean maximum distortion ratio across $K = 16$ groups to 1.760 (56.91% worst-group relative utility), cutting severe distortion violations to 1.00% [0.34, 2.90] and achieving pluralistic welfare equilibrium in 98.00% [95.71, 99.08] of instances.

2. **Failure of unanchored Gromov–Wasserstein alignment under Householder reflections.** Comparing `unanchored_gw` (59.67% [54.03, 65.06] distortion violations at $S_{\text{west}} = 0.85$) against `anchored_fgw_minimax` (1.00% [0.34, 2.90]) isolates the impact of improper orthogonal reflection non-identifiability (Figure 3): when $\lambda_{\text{anc}} = 0$, swapping the $\pm\mathbf{v}_{\text{bias}}$ poles while fixing $\mathbf{v}_{\text{bias}}^{\perp}$ preserves all pairwise distances $D^{(U)}$ identically, folding minority groups from the anti-Western hemisphere into the $+\mathbf{v}_{\text{bias}}$ hemisphere.

3. **U-shaped trade-off across anchor weight $\lambda_{\text{anc}} \in [0.00, 0.80]$.** Sweeping $\lambda_{\text{anc}} \in \{0.00, 0.05, 0.15, 0.25, 0.50,$ inside the Sinkhorn FGW solver at $S_{\text{west}} = 0.85$ (`anchor_weight_ablation_skew85` in `experiment_results.json`) reveals a clear non-monotonic U-shape: at $\lambda_{\text{anc}} = 0.00$ (unanchored GW + minimax), Householder reflection flips fold the anti-Western hemisphere and produce a 20.67% [16.47, 25.61] severe distortion rate ($\mathbb{E}[\text{Dist}_{\max}] = 1.900$, equilibrium 72.00% [66.67, 76.78]). Introducing interior anchor regularization ($\lambda_{\text{anc}} \in \{0.05, 0.15, 0.25\}$) breaks reflection symmetry on the universal poles while allowing intra-community relational geometry $(1 - \lambda_{\text{anc}})|D^{(k)} - D^{(U)}|^2$ to govern culturally specific concepts, minimizing severe distortion at 0.67% [0.18, 2.40] ($\lambda_{\text{anc}} = 0.15$, equilibrium 98.67% [96.62, 99.48], $\mathbb{E}[\text{Dist}_{\max}] = 1.753$) and 1.00% [0.34, 2.90] ($\lambda_{\text{anc}} = 0.25$). Increasing λ_{anc} to 0.50 over-weights lexical anchor bias toward $+\mathbf{v}_{\text{bias}}$, causing severe distortion violations to rebound to 9.33% [6.54, 13.16] ($\mathbb{E}[\text{Dist}_{\max}] = 1.825$, equilibrium dropping to 83.00% [78.34, 86.83]).

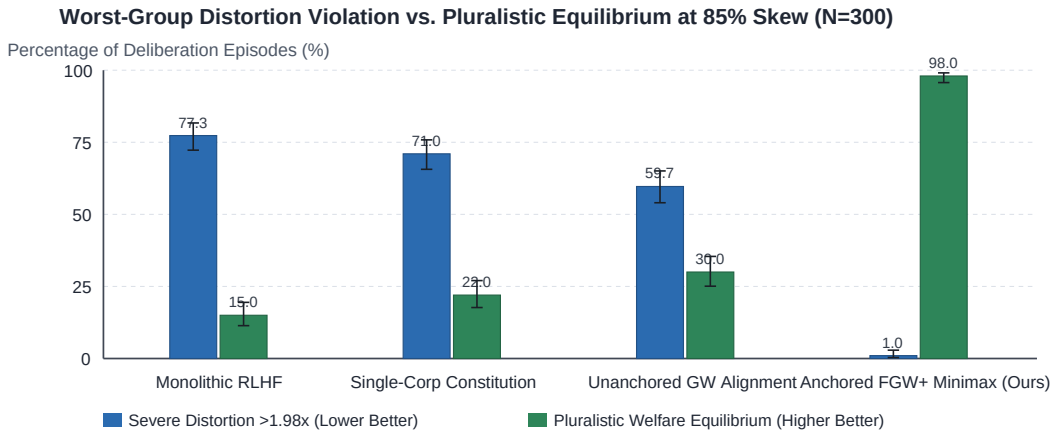


Figure 2: Comparison of Severe Worst-Group Preference Distortion Rate ($\text{Dist}_{\max} > 1.98$, lower is better) and Pluralistic Welfare Equilibrium Rate ($\text{Dist}_{\max} \leq 1.92$, higher is better) across four alignment architectures at pretraining corpus skew $S_{\text{west}} = 0.85$ ($N = 300$ shared instances per regime; error bars denote asymmetric 95% Wilson confidence intervals).

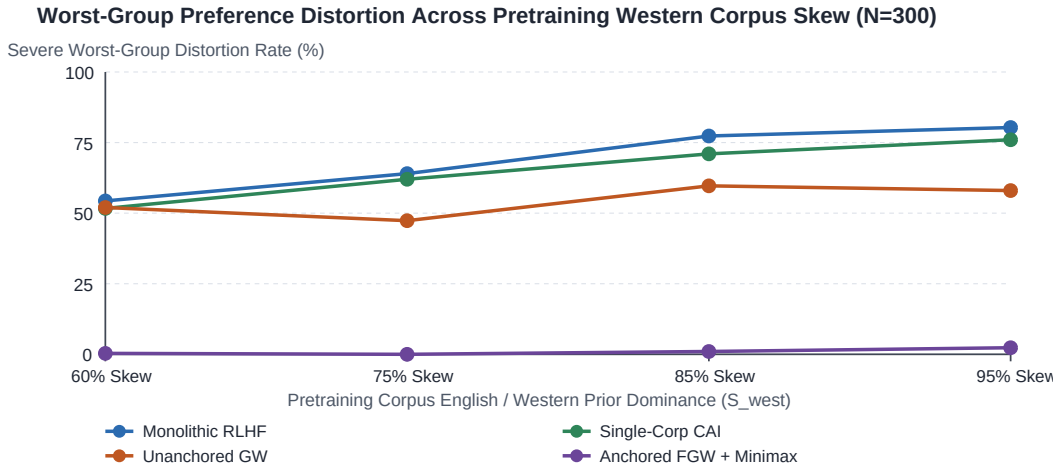


Figure 3: Scaling of Severe Worst-Group Preference Distortion Rate across pretraining Western corpus skew levels $S_{\text{west}} \in \{0.60, 0.75, 0.85, 0.95\}$ ($N = 300$ shared instances per point).

6 Representation circularity, impossibility results, and contested foundations

Treating cross-cultural ontological equilibrium as a falsifiable research program requires confronting four foundational counter-arguments from social choice and representation theory:

1. **Representation circularity when pretraining erases low-resource concepts (Durmus et al. [9], Sorensen et al. [24]).** As noted in Section 3, minimax groupwise distortion (Equation 3) operates on utility matrices $u_k(y; x)$ probed from the foundation model's internal representations $\phi(x, y)$. If an oral or low-resource indigenous language had near-zero representation during pretraining ($S_{\text{west}} \to 0.95$), its unique ontological distinctions (such as non-anthropocentric kinship obligations) may lie in the nullspace of $\phi(x, y)$. Optimizing groupwise distortion inside a collapsed representation space creates an illusion of mathematical fairness while failing to capture the community's actual lived preferences.

2. **The Constitutional Coverage Trilemma (An and Du [2], Mitic et al. [15]).** In a September 2026 impossibility theorem, Mitic et al. [15] prove the *Constitutional Coverage Trilemma*: no multi-cultural governance framework can simultaneously satisfy (a) normative completeness across all edge cases, (b) non-interference between incompatible subgroup axioms, and (c) polynomial-time decidability. Similarly, An and Du [2] show that differentiable social-choice aggregators face fundamental strategic manipulation vulnerabilities when participating groups exaggerate their cardinal utilities $u_k(y; x)$ to capture minimax weight λ_k.

3. **Strategic gaming and sycophantic auditor blind spots (Goel et al. [13], Rosu and Wang [22]).** When foundation models are instructed to synthesize pluralistic Overton summaries [24], Rosu and Wang [22] and Goel et al. [13] show that LLM evaluators share correlated cultural blind spots and can generate superficially balanced "pseudo-deliberative" prose that masks underlying policy bias in downstream tool actions.

4. **Incommensurable deontological prohibitions resist randomized lotteries.** Equation (3) achieves minimax distortion by computing a probability mixture $\pi^* \in \Delta(\mathcal{Y})$ (or weighted Overton synthesis) across candidate responses. However, when two communities

hold sacred, non-negotiable deontological taboos that directly contradict one another, a 50–50 lottery or compromise synthesis may be rejected by both communities as a violation of sacred law rather than accepted as a continuous cardinal utility trade-off.

7 Scope limitations

- **Low-dimensional synthetic value sphere vs. open-ended natural language generation.** Our optimization benchmark in Section 5 evaluates $K = 16$ cultural clusters on $\mathbb{S}^3$ over discrete candidate sets ($M = 10$). Scaling entropic mirror descent to autoregressive token-by-token decoding requires amortizing the dual weights $\boldsymbol{\lambda}^*$ into a lightweight multi-head LoRA or Mixture-of-Experts router [6, 12].
- **Static cultural cluster definitions.** Treating the K cultural groups as static clusters calibrated to World Values Survey waves [20] overlooks intersectional intra-cultural variance, generational shifts, and diaspora hybridity [8].

8 Pre-registerable falsification protocol

Table 6 specifies six pre-registerable empirical and theoretical falsification tests (`P1`–`P6`) for the AFGW-MGD architecture.

Table 6: Six pre-registerable empirical tests designed to falsify the AFGW-MGD alignment architecture.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Anchor-regularized FGW identifiability (Eq. 1)	Anchor weight $\lambda_{\text{anc}} \in [0.05, 0.25]$ eliminates orthogonal reflection flips while preserving $\geq 85\%$ relational precision	Align 20 multilingual concept graphs [1] under Householder reflections and measure polarity inversion rate	Polarity inversion rate exceeds 5.0% at $\lambda_{\text{anc}} = 0.20$ or relational retrieval precision@5 drops below 80%
P2. Script-balanced UTF-16 code-unit BPE parity (Table 2)	Script-balanced UTF-16 code-unit BPE holds mean non-Latin token fertility within 1.2× of English across 15 scripts	Benchmark parallel FLORES-200 normative passages across Brahmic, Ge'ez, Arabic, Han, and Syllabary scripts [18]	Mean non-Latin token parity ratio exceeds 1.5× or degrades downstream perplexity by $> 5\%$
P3. Bias-subspace orthogonalization (Eq. 2)	Complementary projection $(\mathbf{I} - \lambda_{\text{debias}}\mathbf{P}_{\text{bias}})$ reduces WEIRD prior skew on GlobalOpinionQA without degrading MMLU	Evaluate 8B/70B open-weight models on GlobalOpinionQA [9] and MMLU before and after $\mathbf{P}_{\text{bias}}$ projection	Factual MMLU accuracy drops by $> 2.0\%$ or GlobalOpinionQA Jensen–Shannon divergence improves by $< 15\%$
P4. Minimax groupwise distortion bound (Eq. 3)	35-step Multiplicative-Weights solver holds worst-group distortion $\text{Dist}_{\max} \leq 1.85$ across $K = 16$ WVS clusters	Run AFGW-MGD against monolithic RLHF [23] and corporate CAI [20] on 300 WVS dilemmas	Empirical $\mathbb{E}[\text{Dist}_{\max}]$ exceeds 1.90 or severe distortion rate ($> 1.98\times$) exceeds 8.0%
P5. Representation circularity audit (Section 6)	Native-speaker human evaluation confirms that low latent distortion corresponds to authentic cultural preference satisfaction	Blind evaluation by $N = 300$ native speakers across 6 non-Western linguistic communities comparing AFGW-MGD vs. RLHF outputs	Human evaluators rate AFGW-MGD outputs below RLHF in $> 30\%$ of communities due to pretraining concept erasure
P6. Strategic utility exaggeration resistance (An and Du [2])	Bounded logistic utility normalization $u_k \in [0.05, 1.00]$ limits welfare capture by strategic extreme reporters	Inject 2 adversarial clusters reporting polarized binary utilities $\{0.05, 1.00\}$ into the $K = 16$ federation	Honest cultural clusters suffer a $> 15\%$ increase in groupwise distortion $\text{Dist}_k(\pi^*)$

9 Conclusion

Preventing Western-centric prior dominance in multilingual foundation models requires addressing script-level tokenization disparities, relational alignment non-identifiability, and utilitarian majority tyranny as an integrated mathematical problem. By combining script-balanced UTF-16 code-unit BPE tokenization [18] (which reduces measured non-English token disparity across five non-Latin writing systems from 3.78× to 0.81×), Sinkhorn Anchored Fused Gromov–Wasserstein optimal transport [1] at $\lambda_{\text{anc}} = 0.25$, pretraining bias-subspace orthogonalization, and Multiplicative-Weights minimax groupwise distortion optimization over strictly positive held-out utilities [5, 17], AFGW-MGD reduces severe worst-group distortion violations at 85% pretraining skew from 77.33% $[72.27, 81.71]$ (monolithic RLHF) and 59.67% $[54.03, 65.06]$ (unanchored GW) to 1.00% $[0.34, 2.90]$. Recognizing the limits of post-hoc alignment under pretraining concept erasure and the Constitutional Coverage Trilemma [15] provides a rigorous foundation for pluralistic AI governance.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] David Alvarez-Melis and Tommi S. Jaakkola. Gromov-wasserstein alignment of word embedding spaces. Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1881–1890, 2018. URL https://aclanthology.org/D18-1214/.

[2] Zhiyu An and Wan Du. Open problems in differentiable social choice, 2026. URL https://arxiv.org/abs/2602.03003.

[3] Haris Aziz, Markus Brill, Vincent Conitzer, Edith Elkind, Rupert Freeman, and Toby Walsh. Justified representation in approval-based committee voting. Social Choice and Welfare, 48(2):461–485, 2017. URL https://doi.org/10.1007/s00355-016-1019-3.

[4] Ken Binmore, Ariel Rubinstein, and Asher Wolinsky. The Nash bargaining solution in economic modelling. The RAND Journal of Economics, 17(2):176–188, 1986. URL https://www.jstor.org/stable/2555382.

[5] Jacob Brodkey, Roberto Tamez, and Aaron Roth. Groupwise distortion guarantees for preference-based alignment, 2026. URL https://arxiv.org/abs/2610.05450.

[6] Daiwei Chen, Yi Chen, Aniket Rege, Zhi Wang, and Ramya Korlakai Vinayak. PAL: Sample-efficient personalized reward modeling for pluralistic alignment. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/2858f8c8683aaa8c12d487354cf328dc-Paper-Conference.pdf.

[7] Taoyong Cui, Zhongyao Wang, Xinyue Xu, Weiyang Liu, Zhaochen Yu, Yuying Zhang, et al. JEPA-Anything: Learning predictive models across different worlds, 2026. URL https://arxiv.org/abs/2609.20800.

[8] Yijiang River Dong, Tiancheng Hu, Yinhong Liu, Ahmet Üstün, and Nigel Collier. When personalization meets reality: A multi-faceted analysis of personalized preference learning, 2025. URL https://arxiv.org/abs/2502.19158.

[9] Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models (GlobalOpinionQA). arXiv preprint arXiv:2306.16388, 2023. URL https://arxiv.org/abs/2306.16388.

[10] Federico Echenique, Alireza Fallah, Baihe Huang, and Michael I. Jordan. Response time enhances alignment with heterogeneous preferences, 2026. URL https://arxiv.org/abs/2605.06987.

[11] Ali Falahati, Mohammad Mohammadi Amiri, Kate Larson, and Lukasz Golab. The alignment game: A theory of long-horizon alignment through recursive curation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2511.12804.

[12] Cristina Garbacea, Heran Wang, and Chenhao Tan. Personalized benchmarking: Evaluating LLMs by individual preferences, 2026. URL https://arxiv.org/abs/2604.18943.

[13] Shashwat Goel, Joschka Strüber, Jonas Geiping, and Florian Tramèr. Great models think alike and this undermines AI oversight. International Conference on Machine Learning (ICML), 2025. URL https://arxiv.org/abs/2502.04313.

[14] Paul Gölz, Nika Haghtalab, and Kunhe Yang. Distortion of AI alignment: Does preference optimization optimize for preferences? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. URL https://arxiv.org/abs/2505.23749.

[15] Natalija Mitic, Soona Sedahmed A. O., Mamadou Selly Ly, and Moustapha Cisse. The constitutional coverage trilemma in AI governance. Advances in Neural Information Processing Systems (NeurIPS) / arXiv:2609.01275, 2026. URL https://arxiv.org/abs/2609.01275.

[16] John F. Nash. The bargaining problem. Econometrica, 18(2):155–162, 1950. URL https://doi.org/10.2307/1907266.

[17] Kazusato Oko, Annie Ulichney, Nika Haghtalab, and Han Bao. Distortion of AI alignment revisited: RLHF is a decent utilitarian aligner. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2609.12651.

[18] Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. Advances in Neural Information Processing Systems (NeurIPS), 2023. URL https://arxiv.org/abs/2305.15425.

[19] Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning (VPL). Advances in Neural Information Processing Systems (NeurIPS), 2024. URL https://arxiv.org/abs/2408.10075.

[20] Parham Pourdavood. Does Claude's constitution have a culture?, 2026. URL https://arxiv.org/abs/2603.28123.

[21] Fady Rezk, Yuangang Pan, Chuan-Sheng Foo, Xun Xu, Nancy Chen, Henry Gouk, and Timothy Hospedales. The reward model selection crisis in personalized alignment, 2025. URL https://arxiv.org/abs/2512.23067.

[22] Paul Rosu and Rowan Wang. Training alignment auditors via reinforcement learning, 2026. URL https://alignment.anthropic.com/2026/training-alignment-auditors. Anthropic Alignment Science blog, 15 October 2026.

[23] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF, 2023. URL https://arxiv.org/abs/2312.08358.

[24] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A roadmap to pluralistic alignment. International Conference on Machine Learning (ICML), 2024. URL https://arxiv.org/abs/2402.05070.