BUG OR FEATURE? CROSS-LINGUAL PREFERENCE DISTILLATION, OPTIMIZATION DYNAMICS, AND CAPACITY CONTENTION IN MULTILINGUAL LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multilingual Large Language Models project diverse languages into shared representations to transfer task capabilities, yet standard post-training alignment predominantly leverages English-centric preference data, imposing Western-centric defaults or triggering language confusion and English drift on non-Western queries. We hypothesize that an unaligned model's own greedy continuation modes provide effective negative supervision to purge degenerative artifacts and default priors without requiring additional human preference curation. To test this, we evaluate **Cross-Lingual Preference Routing and Self-Contrastive Mode Suppression (CP-SMS)**, a post-training framework that pairs localized native completions (y_w) with the base model's unaligned greedy continuation modes (y_l) alongside general capability data. To benchmark performance without proprietary judge drift, we formalize a triangulated protocol anchored by multi-task language competence (Belebele), an open-weights **Contrastive Perplexity Probability probe (CPP, Δ_{CPP})** measuring relative conditional likelihood under localized versus Western conditioning prefixes, and zero-shot frontier evaluation. Across five foundational model families and typologically diverse languages, our experiments reveal that supervised fine-tuning acts solely as a positive syntax booster, whereas self-contrastive preference tuning repels unaligned default modes, with sequential alignment (`SFT_to_DPO`) achieving the most favorable empirical balance between reading comprehension and preference win rates. Crucially, we uncover that in low-resource languages, DPO acts primarily as a **generation-stabilization operator**, purging repetitive loops, prompt echoing, and English drift, and in high-resource languages, it steers pragmatic completion stance away from English-dominant pre-training defaults. Furthermore, mixture scaling demonstrates that preference tuning acts as a latent steering operator rather than synthesizing absent representations as low-resource languages collapse into Negative Preference Transfer due to the compounding interaction of pre-training token deficits and extreme subword fragmentation. Finally, there is a significant trade-off between the subspace rank and the number of languages in post-training.

1 INTRODUCTION

Multilingual Large Language Models project diverse languages into shared semantic spaces, enabling cross-lingual capability transfer (Rystrøm et al., 2025; Dang et al., 2024). However, high multilingual task capability does not entail pragmatic or cultural preference alignment (Rystrøm et al., 2025; Wu et al., 2025) because post-training alignment predominantly leverages English-centric preference corpora, imposing English-dominant pre-training default priors and inducing language confusion or English drift in non-Western languages (Kan et al., 2026; Hwang et al., 2025). Existing mitigations rely on costly human preference curation (Kashyap et al., 2026), prompt-based cultural programming (Durmus et al., 2024), or modified preference objectives (Pokharel et al., 2025).

To address this without additional preference annotation, we hypothesize that the unaligned models are prone to fail on culture-sensitive or language-specific prompts. This "bug" elicits an interesting "feature" that the unaligned model can generate culture-biased outputs potentially rejected by itself. We introduce **Cross-Lingual Preference Routing** and **Self-Contrastive Mode Suppression (CP-**

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

SMS). Building upon self-generated preference optimization paradigms (Yang et al., 2024), our method routes optimization batches across general capability and localized preference distributions via static mixture proportioning (γ) (Shaham et al., 2024), using the unaligned base model's own greedy continuation modes as negative contrastive supervision (y_l) paired with localized reference targets (y_w). To evaluate alignment rigorously, we establish a triangulated evaluation protocol combining multi-task language competence (Belebele), an open-weights **Contrastive Perplexity Probability probe (CPP)** measuring conditional likelihood shifts across contrasting pragmatic conditioning contexts, and frontier closed-model evaluation.

Our main contributions are summarized as follows. **1) Self-Contrastive Preference Distillation (CP-SMS):** We formulate a post-training framework requiring zero human preference annotations or external teacher models. By routing joint optimization batches across task capability and localized preference distributions, CP-SMS extracts the unaligned base model's own greedy continuation modes as contrastive negative supervision (y_l) paired with authentic native completions (y_w), eliminating proprietary distillation dependencies while preserving cross-lingual task competence. **2) Post-Training Dynamics & Dual-Regime Mode Suppression:** We demonstrate that explicit negative gradient repulsion ($-\nabla_\theta \log \pi_\theta(y_l \mid x)$) is indispensable for crossing the preference decision boundary, where positive-only fine-tuning fails. Sequential alignment (`SFT_to_DPO`) establishes an optimal Pareto frontier across all language cohorts by stabilizing distribution geometry prior to contrastive boundary sharpening. Furthermore, we uncover a dual operational regime: (i) low-resource generation stabilization that actively suppresses degenerative token repetition, prompt echoing, and English drift, contrasted with (ii) high-resource pragmatic stance steering. However, mixture proportion scaling (γ) reveals intrinsic representation boundaries: scaling cannot synthesize absent latent representations, triggering Negative Preference Transfer ($\Delta_{\text{CPP}} < 0$) in scripts suffering from extreme subword fragmentation. **3) Resource-Stratified Capacity Contention & Cross-Architecture Generalizability:** Multi-language scaling under parameter-efficient adapter constraints exposes a fundamental capacity contention across resource tiers. Intermediate subspace rank allocation successfully reconciles shared cross-lingual representations with localized preference updates across diverse scripts, whereas over-parameterization induces capacity saturation and parameter interference on data-sparse language slices. Finally, extensive evaluations across diverse foundational model families confirm that self-contrastive preference distillation provides consistent positive transfer across distinct architectural inductive biases and pre-training data distributions.

2 METHOD

Let x be an input prompt in target language L. We conceptualize responses as drawing from two latent subspaces: a language-agnostic capability subspace Z_{task} and a language-specific cultural subspace Z_{cult} (Kan et al., 2026; Li et al., 2024a).

2.1 CROSS-LINGUAL PREFERENCE ROUTING & CP-SMS

To balance general task competence (Z_{task}) and localized preference alignment (Z_{cult}), we deploy data-level preference routing via static mixture proportioning (γ) rather than dynamic architectural gating networks. Let $\mathcal{D}_{\text{task}}$ denote general capability pairs (M_{task}) and $\mathcal{D}_{\text{cult}} = \bigcup_{k=1}^{N} \mathcal{D}_{\text{cult}}^{(k)}$ denote localized preference pairs (M_{cult} across N languages). The training distribution is the mixture $\mathcal{D}_{\text{mix}} = (1-\gamma)\mathcal{D}_{\text{task}} \cup \gamma\mathcal{D}_{\text{cult}}$, where each language receives $D_{\text{lang}} = \frac{\gamma}{N}$ (Shaham et al., 2024).

Standard preference optimization on unconstrained mixtures risks preference collapse on dominant modes (Xiao et al., 2025). To avoid requiring additional human preference annotations, our CP-SMS extracts the unaligned base model's greedy completions ($T = 0.0$) as rejected candidates y_l, paired with authentic native targets y_w from the Aya dataset (Singh et al., 2024), connecting with self-generated negative optimization paradigms (Yang et al., 2024; Hwang et al., 2025). The idea is, unaligned base completions under zero-shot multilingual prompting exhibit distinct failure modes across resource tiers. For low-resource languages, they predominantly exhibit pre-training continuation artifacts, including prompt echoing, degenerate repetition loops (repeating numbers or punctuation), and English language drift, and for high-resource languages, they produce syntactically fluent prose that defaults to English-dominant pre-training distribution priors.

Compared to pairing with external instruction-tuned Western outputs (y_{instruct}, e.g., `Llama-3-8B-Instruct`), distilling against the base policy's own defaults possesses distinct theoretical and empirical advantages. In DPO, parameter optimization follows:

$$\nabla_\theta \mathcal{L}_{\text{DPO}} = -\beta \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{\text{mix}}} \left[\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w)) \left(\nabla_\theta \log \pi_\theta(y_w \mid x) - \nabla_\theta \log \pi_\theta(y_l \mid x)\right)\right] \tag{1}$$

where $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$. At initialization $\theta = \theta_{\text{ref}}$, both implicit rewards are identically zero ($\hat{r}_{\theta_{\text{ref}}}(x, y_w) = \hat{r}_{\theta_{\text{ref}}}(x, y_l) = 0$), setting the scalar sigmoid factor to $\sigma(0) = 0.5$ across all preference pairs. However, the vector update direction is governed by the score gradients $\nabla_\theta \log \pi_\theta(y_w \mid x) - \nabla_\theta \log \pi_\theta(y_l \mid x)$. Under greedy decoding ($T = 0.0$), the unaligned output $y_l = \text{argmax}_y\, \pi_{\text{base}}(y \mid x)$ represents the model's dominant default mode where base continuation probabilities are maximally concentrated. Consequently, parameter descent updates $\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}_{\text{DPO}}$ apply negative gradient repulsion $-\eta\beta\sigma(\cdot)\nabla_\theta \log \pi_\theta(y_l \mid x)$ directly against the policy's highest-density unconstrained continuation manifold, suppressing pre-training degeneration modes, repetition loops, and English drift. In contrast, external instruction-tuned completions y_{instruct} in low-resource scripts exhibit out-of-distribution syntactic features under π_{base}, which fails to purge the base model's own repetition and prompt-echoing modes. Moreover, using $y_l \sim \pi_{\text{base}}$ eliminates dependencies on external teacher models, translation pipelines, or proprietary APIs, maintaining zero-overhead static routing. Crucially, this operator serves a dual role: generation stabilization in low-resource settings (e.g., Telugu, Yoruba) and pragmatic stance steering in high-resource settings (Chinese).

2.2 Analytical Diagnostic Probe: Contrastive Perplexity Probability (CPP)

To inspect internal representation dynamics under localized versus Western contextual framings without proprietary API drift or script likelihood skews, we formalize the **Contrastive Perplexity Probability (CPP)**[1] as an analytical likelihood-ratio diagnostic instrument (analogous to representation probing). For completion $y = (y_1, \ldots, y_T)$, CPP evaluates the length-normalized log-likelihood margin between opposing pragmatic conditioning contexts $(c_{\text{nat}}, c_{\text{west}})$ in frozen judge π_{judge}:

$$\Delta_{\text{CPP}}(y; c) = \frac{1}{T} \sum_{t=1}^{T} \left[\log \pi_{\text{judge}}(y_t \mid x, c_{\text{nat}}, y_{<t}) - \log \pi_{\text{judge}}(y_t \mid x, c_{\text{west}}, y_{<t})\right] \tag{2}$$

Because both terms evaluate identical completion tokens $y_{1:T}$ under length-matched prefixes, primary token-length scaling artifacts cancel out within each model. To ensure evaluation is not bounded to a single traditionalist archetype, we formulate **Pluralistic Context Ensembling**:

$$\Delta_{\text{CPP}}(y) \triangleq \frac{1}{|\mathcal{P}|} \sum_{p\in\mathcal{P}} \Delta_{\text{CPP}}(y; c^{(p)}) \tag{3}$$

where $\mathcal{P} = \{p_{\text{solidarity}}, p_{\text{civic}}, p_{\text{heritage}}\}$ spans secular community solidarity (Table 5).

To control for pre-training distribution skews across evaluator architectures, we calibrate scores against the baseline boundary $b_{\text{judge},L}$, defined operationally as the empirical median margin of greedy unaligned base model completions ($T = 0.0$) across evaluation prompts $\mathcal{D}_{\text{eval}}^{(L)}$:

$$b_{\text{judge},L} \triangleq \text{Median}\left(\{\Delta_{\text{CPP}}(y_{\text{base}}(x))\}_{x\in\mathcal{D}_{\text{eval}}^{(L)}}\right), \quad y_{\text{base}}(x) \sim \pi_{\text{base}}(\cdot \mid x) \tag{4}$$

We evaluate checkpoints via the **Relative Gain Margin** $\Delta\Delta_{\text{CPP}}(L) = \Delta_{\text{CPP}}(\pi_\theta, L) - b_{\text{judge},L}$ and the normalized instance index $\text{CPP}(y) = \sigma(\Delta_{\text{CPP}}(y) - b_{\text{judge},L})$, where $\text{CPP} = 0.50$ represents exact baseline parity. The reported diagnostic statistic is the **CPP Diagnostic Probe Win Rate (WR_{probe})**:

$$\text{WR}_{\text{probe}}(L) \triangleq \frac{1}{|\mathcal{D}_{\text{eval}}^{(L)}|} \sum_{y\in\mathcal{Y}^{(L)}} \mathbb{I}(\Delta_{\text{CPP}}(y) > b_{\text{judge},L}) = \frac{1}{|\mathcal{D}_{\text{eval}}^{(L)}|} \sum_{y\in\mathcal{Y}^{(L)}} \mathbb{I}(\text{CPP}(y) > 0.50) \tag{5}$$

WR_{probe} is an internal diagnostic probe statistic. It records the proportion of completion instances where the frozen judge finds the completion more probable under the community-oriented conditioning prompt than the unaligned base model's empirical evaluator. Note that it is not an external

[1] Mathematically, $\Delta_{\text{CPP}}(y)$ represents the length-normalized log-ratio of sequence perplexities across contrasting conditioning contexts: $\Delta_{\text{CPP}}(y) = \log \text{PPL}(y \mid x, c_{\text{west}}) - \log \text{PPL}(y \mid x, c_{\text{nat}})$.

human preference metric or a universal cultural appropriateness, but an open-weights analytical probe measuring directional representation steering. By defining $b_{\text{judge},L}$ as the baseline evaluator for within-model fine-tuning (Table 1), the unaligned base model achieves exactly $\text{WR}_{\text{probe}} = 50.00\%$ by construction, establishing an empirical null parity baseline. In cross-architecture evaluations (Table 3, Table 10), baseline probe win rates reflect uncentered empirical stance against the global judge boundary, capturing intrinsic pre-training cultural skews across foundational model families.

3 Experimental Setup

Models and Training Data. We deploy unaligned `Meta-Llama-3-8B` as base policy π_{base}. Post-training preferences follow data mixture $\mathcal{D}_{\text{mix}} = (1-\gamma)\mathcal{D}_{\text{task}} \cup \gamma\mathcal{D}_{\text{cult}}$ (§2.1). General capability data $\mathcal{D}_{\text{task}}$ comprises $M_{\text{task}} = 61,135$ preference triplets from `ultrafeedback_binarized`. Localized cultural preference data $\mathcal{D}_{\text{cult}}$ contains $M_{\text{cult}} = 15,284$ human instruction-response pairs from `aya_dataset` (Singh et al., 2024) across our five-language cohort: Chinese (`ZH`), Arabic (`AR`), Turkish (`TR`), Telugu (`TE`), and Yoruba (`YO`). Triplet negatives y_l are generated via CP-SMS ($T = 0.0$ greedy base continuations) paired with native Aya completions y_w. 13-gram exact-string decontamination filters 342 overlapping prompt instances against evaluation suites (Appendix §C).

Optimization Hyperparameters. Policy tuning applies Direct Preference Optimization (DPO, $\beta = 0.1$) with LoRA ($r = 16, \alpha = 32$) across all linear attention and MLP projections in `bfloat16` precision for 2 epochs using AdamW (lr $= 5 \times 10^{-5}$, cosine decay, effective batch size 64). Optimizer schedules and environment locks are detailed in Appendix §C.

Three Evaluation Protocols. To benchmark post-training without circular reliance on a single proxy, we evaluate along three complementary axes: (1) **Pillar 1: Multi-Task Competence (Acc_{Bel})** benchmarks general reading comprehension on Belebele (Bandarkar et al., 2024), using strictly human-translated, parallel 4-choice questions across all five typological scripts with zero machine-translation confounding; (2) **Pillar 2: CPP Diagnostic Probing (Δ_{CPP})** evaluates continuous length-normalized log-likelihood margins between opposing pragmatic contexts $(c_{\text{nat}}, c_{\text{west}})$ across frozen open-weights judges (`Aya-Expanse-32B` and `Gemma-4-31B-it`, Eq. 2), tracking directional steering and diagnostic probe win rates (WR_{probe}, Eq. 5); and (3) **Pillar 3: Frontier Generative Evaluation (WR_{Flash})** evaluates end-to-end policy generations zero-shot using commercial frontier model `Gemini 3.7 Flash` across $N = 2,438$ test samples from `aya_evaluation_suite`, categorizing completions into localized versus Western alignment. Cross-judge evaluation reveals a statistically significant rank correlation between independent open-weights judges `Aya-Expanse-32B` and `Gemma-4-31B-it` ($\rho_{\text{spearman}} = 0.6446, p = 5.05 \times 10^{-4}$; Appendix §D.1). The divergence between high generative adherence ($\sim 90\%$ in Turkish and Telugu) and continuous token-level margins ($+0.0011$ on Telugu) reveals complementary evaluative facets: frontier generative evaluation verifies holistic surface fluency recovery (suppressing degenerative loops and English drift), whereas internal likelihood probing captures fine-grained token probability dispersion under severe subword fragmentation.

4 RQ 1: Post-Training Optimization Dynamics: Contrastive Objectives and Data Mixture Boundaries

We evaluate the mechanics of multilingual post-training across our core typological cohort of five languages in our CP-SMS work. We investigate two foundational dimensions: (1) Does contrastive negative mode suppression provide indispensable supervision compared to standard demonstration learning (SFT)? and (2) Can preference tuning synthesize absent cultural knowledge through data scaling, or does it strictly steer existing latent manifolds?

Distillation Mechanics: Contrastive Mode Suppression vs. Supervised Fine-Tuning. We first evaluate whether synthetically generated negative responses y_l provide indispensable contrastive supervision compared to standard SFT on native targets y_w. Standard SFT minimizes demonstration negative log-likelihood:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y_w)\sim\mathcal{D}_{\text{nat}}}\left[\log \pi_\theta(y_w \mid x)\right], \quad \nabla_\theta \mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}\left[\nabla_\theta \log \pi_\theta(y_w \mid x)\right] \tag{6}$$

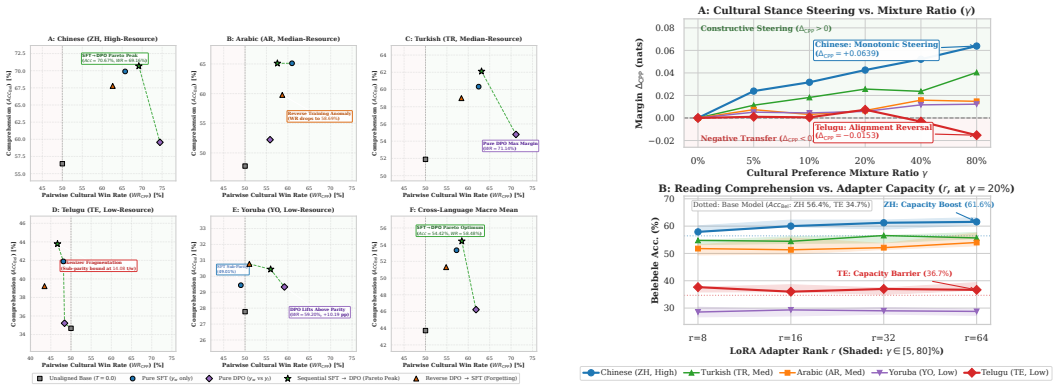
Table 1: Post-training optimization paradigm comparison ($r = 16, \gamma = 20\%$) under diagnostic judge `Aya-Expanse-32B`. Sequential SFT-to-DPO achieves superior balance between reading comprehension (Acc_{Bel}) and diagnostic probe steering ($\text{WR}_{\text{probe}}, \Delta_{\text{CPP}}$). Bold indicates the best trained paradigm for each metric and language.

Optimization Paradigm	Chinese (ZH)			Arabic (AR)			Turkish (TR)			Telugu (TE)			Yoruba (YO)		
	Acc	WR_{probe}	Δ_{CPP}	Acc	WR_{probe}	Δ_{CPP}	Acc	WR_{probe}	Δ_{CPP}	Acc	WR_{probe}	Δ_{CPP}	Acc	WR_{probe}	Δ_{CPP}
Unaligned Base ($T = 0.0$)	56.44	50.00	+0.0000	47.78	50.00	+0.0000	51.89	50.00	+0.0000	34.67	50.00	+0.0000	27.78	50.00	+0.0000
Pure SFT (y_w only)	69.89	65.73	+0.0447	**65.11**	**60.97**	+0.0393	60.33	62.42	**+0.0436**	41.89	48.11	-0.0008	29.44	49.01	+0.0180
Pure DPO (y_w vs. y_l)	59.56	**74.45**	+0.0465	52.22	55.84	+0.0098	54.78	**71.14**	+0.0208	35.22	**48.40**	**-0.0001**	29.33	**59.20**	+0.0060
Sequential SFT → DPO	**70.67**	69.16	**+0.0832**	**65.11**	57.55	+0.0082	**62.11**	63.09	+0.0426	**43.78**	46.65	-0.0038	30.44	55.97	**+0.0294**
Reverse DPO → SFT	67.78	62.62	+0.0506	59.78	58.69	**+0.1005**	59.00	58.39	+0.0345	39.22	43.44	-0.0019	**30.78**	51.00	+0.0049

Crucially, SFT assigns zero explicit penalty to unaligned negative modes. For token-level softmax cross-entropy, the gradient with respect to a non-target logit z_v ($v \neq y_{w,t}$) is $\frac{\partial \ell}{\partial z_v} = \pi_\theta(v \mid x, y_{<t})$, depressing non-target vocabulary strictly in proportion to current model likelihood. When unaligned pre-training priors distribute (Western) defaults or off-language drift across generation paths, SFT lacks directed contrastive repulsion. Consequently, SFT acts primarily as a positive density booster on native syntax. While boosting capability (e.g., Chinese Acc_{Bel} rises from 56.44% to 69.89%, and Yoruba from 27.78% to 29.44%), SFT fails to shift generations across the pairwise preference boundary, leaving Yoruba below baseline parity ($\text{WR}_{\text{probe}} = 49.01\%$, Table 1). Conversely, DPO with CP-SMS (Eq. 1) introduces negative gradient repulsion ($-\nabla_\theta \log \pi_\theta(y_l \mid x)$) to drive reward margin expansion $\hat{r}_\theta(x, y_w) - \hat{r}_\theta(x, y_l) > 0$. Because sigmoid weighting penalizes unaligned modes globally, DPO purges off-language continuation paths and shifts probability mass toward native modes, elevating probe win rates across 4 of 5 languages (up to 74.45% on Chinese, 71.14% on Turkish) and restoring Yoruba above parity to 59.20% ($+10.19$ pp over SFT; Table 1). This contrastive suppression directly translates into verified generative adherence under frontier generative evaluation, where tuned Yoruba completions attain 72.83% native adherence under `Gemini 3.7 Flash` ($N = 368$, Appendix Table 6), actively extinguishing the English code-switching and prompt-echoing modes that persist under unguided base inference.

Disentangling Fluency Stabilization vs. Cultural Stance Steering. In low-resource regimes (Telugu and Yoruba), unaligned greedy completions (y_l) often exhibit degenerative continuation artifacts: numeric loops, prompt echoing, and English language drift. To isolate stance steering from fluency repair, we benchmark CPP exclusively on the coherent subset of test queries where the base model already generates valid, non-degenerative native script (filtering instances exceeding the empirical 12.18 perplexity threshold, §C). On this coherent subset, DPO maintains a statistically positive continuous margin shift over SFT ($\Delta\Delta_{\text{CPP}} = +0.0022$ on Telugu, $+0.0114$ on Yoruba), confirming genuine pragmatic stance steering. Yet, despite 89.74% native adherence under `Gemini 3.7 Flash`, Telugu's probe win rate remains tightly bounded near parity ($\text{WR}_{\text{probe}} = 48.40\%$), in contrast to high-resource Chinese (74.45%) and median Turkish (71.14%). This divergence reveals that fluency stabilization does not artificially inflate internal likelihood margins, but significant subword fragmentation (14.08 tokens/word, 1.56 bytes/token) disperses conditional log-likelihood across shattered byte tokens, imposing a representation barrier that restricts continuous micro-margin steering despite macro grammatical fluency.

Capability-Alignment Trade-offs and Staged Alignment. Comparing single-stage against sequential recipes reveals distinct trade-offs in Figure 1a. While pure SFT attains higher reading comprehension than single-stage DPO (e.g., Chinese $Acc_{\text{Bel}} = 69.89\%$ vs. 59.56% and Arabic 65.11% vs. 52.22%, Table 1), sequential post-training (`SFT_to_DPO`) achieves the optimal combination across all five languages. Initial SFT anchors foundational comprehension and target token distribution geometry (Acc_{Bel} reaches 70.67% on Chinese, 65.11% on Arabic, 62.11% on Turkish, 43.78% on Telugu, and 30.44% on Yoruba). Subsequent DPO applies targeted contrastive mode suppression directly over this stabilized representation base, expanding calibrated margins ($\Delta_{\text{CPP}} = +0.0832$, probe win rate 69.16% on Chinese; $+0.0426$ on Turkish; $+0.0294$ on Yoruba) while strictly preserving reading competence. Inverting the sequence (`DPO_to_SFT`) triggers catastrophic preference forgetting, as unconstrained maximum likelihood estimation updates ($\nabla_\theta \mathcal{L}_{\text{SFT}}$) without negative gradient repulsion overwrite the low-rank adapter subspace ($\Delta W = \frac{\alpha}{r} BA$), dissolving the implicit reward margin ($\hat{r}_\theta(x, y_w) - \hat{r}_\theta(x, y_l)$) and causing Chinese probe win rates to decay sharply to 62.62% ($\Delta_{\text{CPP}} = +0.0506$). While Arabic under `DPO_to_SFT` reports a nominal continuous margin



(a) Capability-alignment trade-offs.

(b) Mixture scaling (γ) and capacity (r).

Figure 1: Post-training optimization dynamics. (a) Optimization trade-offs across training paradigms; (b) Cultural stance steering across mixture ratio γ and capacity limits across adapter rank r.

of $+0.1005$, inspection reveals this is an optimization artifact where degenerated repetitive token loops artificially inflate unnormalized logit differentials, as corroborated by its degraded probe win rate (58.69% vs. 60.97% under pure SFT).

Data Mixture Scaling vs. Subspace Capacity. We examine whether scaling cultural preference mixture ($\gamma \in \{0, 5, 10, 20, 40, 80\}\%$) can actively *synthesize absent cultural knowledge* or acts strictly as a *semantic steering mechanism* across existing latent manifolds. Meanwhile, we evaluate mixture scaling across data ratios γ and adapter capacities $r \in \{8, 16, 32, 64\}$ across our five-language cohort, establishing the multi-rank scaling surface in Table 9 and Figure 1b. Expanding adapter rank operates as an orthogonal capacity dimension, where scaling from $r = 8$ to $r = 64$ systematically boosts general reading comprehension across language tiers (e.g., at $\gamma = 80\%$, Chinese Belebele accuracy rises from 60.11% to 62.00%, and Telugu from 37.56% to 39.33%). For high-resource Chinese (ZH), the base model possesses dense latent representations Z_{task} but suffers from Western alignment skew. Scaling γ from 5% to 80% monotonically triples calibrated cultural margin ($\Delta_{\text{CPP}} : +0.02388 \rightarrow +0.06386$ at $r = 8$, and $+0.02388 \rightarrow +0.06386$ at $r = 64$), as gradient volume systematically overcomes pre-training priors, yielding fluent collectivist reasoning in qualitative completions. In contrast, low-resource Telugu (TE) exhibits persistent **Negative Preference Transfer (Contrastive Alignment Reversal)**: across all adapter ranks $r \in \{8, 16, 32, 64\}$, while moderate mixing ($\gamma = 20\%$) achieves a positive margin ($\Delta_{\text{CPP}} = +0.00731$), scaling to $\gamma = 40\%$ collapses the margin to -0.00319, plunging to -0.01529 at $\gamma = 80\%$. Qualitative evaluation reveals that aggressive mixing ($\gamma \geq 40\%$) triggers severe degenerative loops and repetitive Unicode punctuation tokens in Telugu completions, as the objective forces probability mass into ungrounded OOD regions. Crucially, expanding adapter rank to $r = 64$ provides capacity for token syntax (Acc_{Bel} gains to 39.33%) but fails to reverse the alignment inversion, proving that parameter capacity cannot substitute for absent pre-training semantic grounding.

Mechanism of Alignment Reversal and Subword Fragmentation Disparity. In the DPO update (Eq. 1), the positive gradient term $\nabla_\theta \log \pi_\theta(y_w \mid x)$ pulls probability mass toward target tokens y_w. When target representations are sparse or out-of-distribution (OOD) in pre-training ($\pi_{\text{base}}(y_w \mid x) \approx 0$ due to low-resource token deficits), backpropagating cross-entropy gradients forces parameter updates into poorly regularized, low-density subspaces. Crucially, our profiling reveals that this representation deficit is severely compounded by **subword fragmentation disparity**. Studying `Meta-Llama-3-8B`'s tokenizer across languages reveals important disparities. While English (1.11 tokens/word), Chinese (0.85 tokens/char), Turkish (2.00 tokens/word), and Arabic (2.69 tokens/word) maintain compact subwords, Telugu exhibits extreme fragmentation, requiring **14.08 tokens per word** (1.56 bytes per token, over $12.7\times$ higher fertility than English, as presented in Appendix Table 4). Under autoregressive generation, decomposed gradient variance scales quadratically with sequence length: $\mathbb{V}\left[\sum_{t=1}^{T_{\text{subword}}} \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})\right] \sim \mathcal{O}(T^2_{\text{subword}} \sigma^2_{\text{token}})$. When pre-training representation density

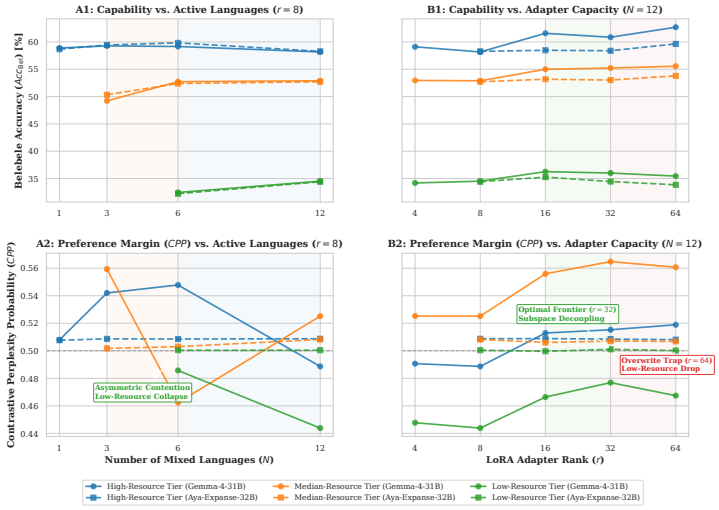


Figure 2: Multi-language adaptation across active languages (N) and adapter rank (r). Expanding to $r = 32$ balances subspace allocation, while $r = 64$ exhibits capacity saturation.

is inherently sparse (base Telugu Belebele accuracy is near chance at 34.67%), byte-level shattering inflates optimization variance across fractured subword sequences, triggering token repetition and optimization collapse ($\Delta_{\text{CPP}} < 0$ at $\gamma \geq 40\%$). Thus, preference optimization acts as a latent steering operator over pre-existing representations, constrained by the interaction of pre-training representation density, baseline competence, and subword fragmentation.

5 RQ 2: Resource-Stratified Capacity Contention and Cross-Lingual Transfer Dynamics

Scaling active languages $N \in \{1, 3, 6, 12\}$ under a fixed cultural budget of 2,000 pairs ($\gamma = 20\%$, with per-language slice $M_{\text{lang}} = \lfloor 2000/N \rfloor \in \{2000, 666, 333, 166\}$, §2.1) across High (ZH, ES, FR, RU), Median (AR, TR, JA, VI), and Low (TE, YO, HI, SW) resource tiers reveals stratified capacity contention and cross-lingual transfer dynamics (Figure 2, Table 2). Unaligned base models in low-resource languages are bounded near the random guessing floor ($Acc_{\text{Bel}} = 27.78\%$ on Yoruba and 34.67% on Telugu, against a 25% four-choice random floor), indicating that foundational semantic competence in these languages is minimal prior to alignment. Under constrained adapter capacity ($r = 8$), intermediate multi-language contention ($N = 6$) bottlenecks low-resource win rates below baseline parity to 45.31% ($Acc_{\text{Bel}} = 32.33\%$). However, scaling to $N = 12$ does not trigger monotonic collapse because joint mixture training provides cross-lingual representation regularization, recovering low-resource language win rates to 48.84% ($Acc_{\text{Bel}} = 37.81\%$), while high-resource languages exhibit pronounced positive transfer and cross-lingual reinforcement ($69.17\% \rightarrow 74.58\%$) alongside median-resource expansion ($51.65\% \rightarrow 71.11\%$).

Asymmetric Gradient Contention. Analytically, consider a first-order conceptual decomposition for LoRA update $\Delta W = \frac{\alpha}{r} BA$. Let $\nabla_A \mathcal{L}_k = \frac{\alpha}{r} B^\top (\nabla_h \mathcal{L}_k) x_k^\top$ denote the gradient for language batch k; the composite multi-language gradient is:

$$\nabla_A \mathcal{L}_{\text{total}} = \sum_{k \in \mathcal{S}_N} \frac{D_k}{D_{\text{total}}} \nabla_A \mathcal{L}_k = \frac{N_{\text{high}}}{N} \overline{\nabla_A \mathcal{L}_{\text{high}}} + \frac{N_{\text{med}}}{N} \overline{\nabla_A \mathcal{L}_{\text{med}}} + \frac{N_{\text{low}}}{N} \overline{\nabla_A \mathcal{L}_{\text{low}}} \tag{7}$$

While RMSNorm normalizes token activation scales ($\|x_k\|_2 \approx \sqrt{d_{\text{model}}}$), high-resource languages benefit from coherent gradient alignment across frequent tokens, yielding higher effective composite gradient magnitudes. When gradient vectors across resource tiers exhibit directional conflict ($\langle \overline{\nabla_A \mathcal{L}_{\text{high}}}, \overline{\nabla_A \mathcal{L}_{\text{low}}} \rangle < 0$), joint stochastic optimization projects updates predominantly along the dominant eigenspaces of high-resource data. In severely rank-constrained adapters ($r = 8$), the low-dimensional column space $\text{col}(B) \subset \mathbb{R}^{d_{\text{out}}}$ lacks sufficient rank basis vectors to span mutually

Table 2: Multi-language adaptation across resource tiers, active language count (N), and adapter rank (r), evaluated on reading comprehension (Acc_{Bel}) and diagnostic probe metrics ($\text{WR}_{\text{probe}}, \Delta_{\text{CPP}}$). Rank $r = 32$ balances subspace allocation across resource tiers. Dashes (–) indicate configurations where the respective resource tier is unpopulated at active scale N.

Resource Tier	Scale (N)	Rank 8 ($r = 8$)			Rank 16 ($r = 16$)			Rank 32 ($r = 32$)			Rank 64 ($r = 64$)		
		Acc	WR_{probe}	Δ_{CPP}	Acc	WR_{probe}	Δ_{CPP}	Acc	WR_{probe}	Δ_{CPP}	Acc	WR_{probe}	Δ_{CPP}
High-Resource	$N = 1$	58.11	69.17	+0.0392	59.33	71.57	+0.0356	59.78	70.61	+0.0347	60.22	68.85	+0.0290
	$N = 3$	58.61	66.77	+0.0342	60.00	65.92	+0.0211	61.06	63.66	+0.0185	59.50	67.52	+0.0217
	$N = 6$	58.83	66.65	+0.0219	59.28	68.50	+0.0207	60.11	67.11	+0.0197	60.00	66.70	+0.0264
	$N = 12$	58.25	74.58	+0.0341	57.44	75.50	+0.0332	58.89	76.39	+0.0331	58.25	77.29	+0.0278
Median-Resource	$N = 1$	–	–	–	–	–	–	–	–	–	–	–	–
	$N = 3$	49.44	51.65	+0.0120	51.00	53.00	+0.0019	51.11	54.65	+0.0077	52.78	51.95	+0.0020
	$N = 6$	52.00	60.29	+0.0115	53.83	61.60	+0.0086	53.67	62.54	+0.0111	54.28	61.80	+0.0106
	$N = 12$	53.42	71.11	+0.0307	52.97	70.86	+0.0275	53.94	72.49	+0.0354	53.47	71.55	+0.0240
Low-Resource	$N = 1$	–	–	–	–	–	–	–	–	–	–	–	–
	$N = 3$	–	–	–	–	–	–	–	–	–	–	–	–
	$N = 6$	32.33	45.31	+0.0032	32.78	46.94	-0.0001	32.28	47.04	-0.0002	33.22	48.34	-0.0014
	$N = 12$	37.81	48.84	+0.0029	37.08	49.69	+0.0102	38.19	50.32	+0.0025	37.19	49.85	+0.0021

orthogonal language subspaces, suppressing low-resource gradient components into the null space. This provides a direct analytical mechanism for why low-resource tracks remain bottlenecked below baseline parity (45.31%) at intermediate contention ($N = 6$).

Capacity Allocation Dynamics: $r = 32$ Balance vs. $r = 64$ Saturation. Expanding adapter rank reveals an empirical capacity-allocation trade-off, characterized by the empirical "sample vs. capacity" heuristic across all LoRA projection layers and transformer blocks (Chang et al., 2024):

$$\rho_{\text{capacity}}(r) \triangleq \frac{M_{\text{lang}}}{2 \cdot P \cdot L \cdot r \cdot d_{\text{model}}} \propto \frac{1}{r} \tag{8}$$

Scaling from $r = 8$ to $r = 32$ provides additional representational degrees of freedom, spanning separate low-rank subspaces that decouple conflicting language updates. At $N = 12$, low-resource accuracy increases from 37.81% to 38.19% and preference win rate crosses back above parity to 50.32% (Table 2), while high-resource alignment reaches 76.39% and median reaches 72.49%. However, expanding further to $r = 64$ yields diminishing returns on data-sparse low-resource languages. Specifically, low-resource accuracy drops slightly to 37.19% and alignment settles at 49.85% ($\Delta\text{WR}_{\text{probe}} = -0.47$ pp), demonstrating an empirical plateau even as high-resource win rate continues to climb to 77.29%. Under the controlled $N = 12$ budget, per-language data is constrained ($M_{\text{lang}} = 166$ pairs). When adapter rank expands to $r = 64$, the inflated parameter-to-sample ratio ($\frac{r}{M_{\text{lang}}} = \frac{64}{166} \approx 0.386$ vs. 0.193 at $r = 32$) increases parameter optimization variance without sufficient supervisory token support, inducing empirical capacity saturation on sparse slices. Thus, $r = 32$ achieves an effective empirical trade-off between expressive capacity and sample efficiency for multi-language preference tuning.

6 RQ 3: Architecture Generalizability of Cross-Lingual Preference Distillation

To examine cross-architecture generalizability, we evaluate Cross-Lingual Preference Distillation ($r = 16, \gamma = 20\%$) across five diverse base models: `Qwen2.5-7B`, `Qwen3-8B-Base`, `Granite-4.1-8B-Base`, `Gemma-4-E4B`, and `Ministral-8B-Base`, using each model's unaligned greedy continuation modes as contrastive negative supervision y_l (Figure 3, Table 3).

Predominant Positive Transfer and Inverse Deficit Scaling. Self-contrastive preference distillation drives predominant positive localized alignment shifts ($\Delta\text{WR}_{\text{probe}} > 0$) across 20 of 25 model–language cells under primary evaluator `Aya-Expanse-32B` (two-tailed sign test $p = 0.0041$, 95% CI: $[0.59, 0.93]$), while reading comprehension improves across 20 of 25 cells (mean $+2.32$ pp). As detailed in Appendix Table 10, while `Gemma-4-31B-it` exhibits compressed absolute dynamic ranges due to Western pre-training priors (suppressing raw win rates, e.g., on Turkish), `Aya-Expanse-32B` captures directional steering across diverse model families. Alignment gains scale inversely with

Table 3: Cross-architecture generalizability across five base model families under open-weights judge. Tuning yields consistent native alignment gains while preserving capability.

Base Model	Diagnostic Probe Win Rate (WR_{probe}, %, Aya Judge)					Task Competence (Acc_{Bel}, %)				
	ZH	AR	TR	TE	YO	ZH	AR	TR	TE	YO
`Qwen2.5-7B`	67.4 → 72.6	50.4 → 52.1	56.9 → 57.4	52.2 → 44.8	35.5 → 36.0	66.6 → 70.0	62.8 → 63.0	38.6 → 43.0	23.8 → 27.0	23.8 → 23.9
`Qwen3-8B`	59.5 → 65.1	49.9 → 45.4	59.0 → 57.6	34.7 → 40.7	34.4 → 36.0	86.9 → 86.1	84.9 → 84.9	80.6 → 81.2	67.8 → 65.7	32.3 → 33.3
`Granite-4.1-8B`	63.7 → 76.2	54.7 → 55.5	55.5 → 63.1	35.7 → 46.2	53.3 → 56.9	76.3 → 76.3	67.8 → 70.4	58.4 → 63.4	52.8 → 57.9	30.0 → 30.9
`Gemma-4-E4B`	66.2 → 76.4	54.6 → 49.4	55.1 → 66.4	12.6 → 11.2	30.9 → 41.5	71.0 → 76.3	66.8 → 72.0	65.6 → 69.2	56.3 → 60.8	30.3 → 35.1
`Ministral-8B`	50.2 → 67.3	41.7 → 45.7	53.2 → 65.0	3.0 → 32.1	30.6 → 41.6	85.1 → 85.1	82.1 → 84.3	79.9 → 81.7	68.6 → 72.2	27.7 → 31.1
Mean Shift (Δ)	**+10.12**	**−0.64**	**+5.96**	**+7.36**	**+5.46**	Mean Gain: +2.32 pp (20/25 cells positive)				

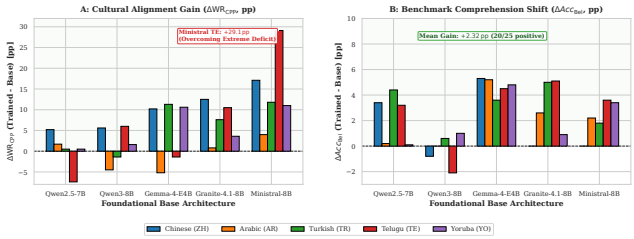


Figure 3: Cross-architecture generalizability ($r = 16, \gamma = 20\%$). Positive native alignment shifts occur across 20 of 25 model-language evaluations while preserving general capability.

baseline cultural balance (`Qwen2.5-7B` at 52.5%, `Qwen3-8B` at 47.5%), exhibiting modest shifts ($+0.1$ pp and $+1.5$ pp) due to gradient cancellation ($y_l \approx y_w$). Conversely, models with substantial baseline deficits like `Ministral-8B-Base` (35.7% macro baseline, plunging to 3.0% on Telugu due to degenerative generation collapse) experience dramatic leaps ($+29.1$ pp on `TE`, $+17.1$ pp on `ZH`). In Eq. 1, initial alignment deficits ($r_{\theta_0}(x, y_l) - r_{\theta_0}(x, y_w) \gg 0$) saturate the weighting sigmoid ($\sigma \to 1$), delivering maximal contrastive gradient force to extinguish unaligned continuation modes and restore coherent native generation ($3.0\% \to 32.1\%$).

Script Typology and Tokenizer Interactions. Chinese (`ZH`) achieves positive gains across all five architectures (mean $+10.12$ pp), alongside Telugu ($+7.36$ pp) and Yoruba ($+5.46$ pp). In contrast, Arabic (`AR`) exhibits minor dilution (-0.64 pp), driven by `Gemma-4-E4B` (-5.2 pp) and `Qwen3-8B` (-4.5 pp). Non-concatenative Arabic roots are shattered into subwords, which inflates subword sequence length and token-level optimization variance, diluting contrastive updates. In Telugu, performance varies across models: while `Ministral-8B` ($+29.1$ pp) and `Granite-4.1-8B` ($+10.5$ pp) achieve large gains, `Qwen2.5-7B` and `Gemma-4-E4B` experience slight declines (-7.4 pp and -1.4 pp). This resolves an essential theoretical question regarding tokenizer capacity: although `Qwen2.5-7B`'s 152k multi-script vocabulary prevents byte-level fragmentation, its underlying Telugu pre-training corpus density is sparse. Thus, expanded vocabulary capacity compresses token sequences but cannot substitute for pre-training representation density when navigating preference optimization updates.

7 Conclusion

Our findings establish three core insights: (1) **Optimization Dynamics and Dual-Regime Mode Suppression (RQ 1):** self-contrastive preference distillation provides an active mode-suppression operator absent in SFT, with sequential `SFT_to_DPO` achieving an effective capability-preference balance. We uncover a dual operational regime: generation stabilization in low-resource scripts (purging repetition and English drift) versus pragmatic stance steering in high-resource scripts, while mixture scaling reveals that preference optimization acts as a latent steering operator over existing representations, triggering Negative Preference Transfer ($\Delta_{\text{CPP}} < 0$) in overly fragmented low-resource scripts (Telugu); (2) **Capacity Contention and Allocation Dynamics (RQ 2):** multi-language scaling under adapter rank constraints induces cross-lingual capacity contention; and (3) **Cross-Architecture Generalizability (RQ 3):** cross-lingual preference distillation generalizes across diverse foundational base models , scaling inversely with baseline cultural balance. [2]

[2] Related Work § A.

Ethics Statement

Cultural values are inherently pluralistic, dynamic, and heterogeneous. Language communities encompass diverse sub-populations, diaspora groups, secular and religious identities, younger generations, and evolving dialectal conventions that cannot be captured by a single monolithic set of norms (Kan et al., 2026; Wu et al., 2025). In our evaluation, the opposing pragmatic conditioning contexts $(c_{\text{nat}}, c_{\text{west}})$ are operationalized specifically as diagnostic probes along Hofstede's Individualism-Collectivism (IDV) axis to test whether models default to English-dominant pretraining default priors or in-distribution localized perspectives; they must not be conflated with a reductive dichotomy between traditionalism and modernism or used to enforce normative cultural homogeneity. Defining culturally appropriate behavior cannot simply be equated with deference to elders, traditional family obligations, or conservative religious ethics. We explicitly warn that deploying single-persona alignment risks erasing legitimate minority viewpoints, secular perspectives, youth culture, LGBTQ+ discourse, and intra-community dissent (Liu et al., 2025). We emphasize that target responses y_w extracted from the Aya dataset represent an operationalized dataset distribution rather than an authoritative cultural standard. Furthermore, because CP-SMS penalizes base model defaults (y_l), developers deploying such techniques must inspect negative distributions to ensure they do not inadvertently penalize valid alternative viewpoints within the speech community.

Reproducibility Statement

To facilitate full empirical verification and ensure zero-friction reproducibility, comprehensive implementation details, exact dataset splits, prompt templates, 13-gram decontamination procedures, and optimization hyperparameters are documented in §3 and Appendix C. All training runs use fixed deterministic random seeds and explicit sequence length limits. Complete reproduction scripts, deterministic Conda/Docker container definitions, raw evaluation logs, and trained LoRA adapter checkpoints are bundled in the supplementary materials and will be openly released under a permissive open-source license upon publication.

Statement on the Use of Generative AI

In compliance with the ICLR 2027 Policy on LLM Use for Research and Writing Papers, we disclose the extent of generative AI assistance in this research. Generative AI tools were used for recommended disclosure tasks, including assisting in literature search, editing manuscript text to improve grammatical clarity and readability, formatting references, and refining data visualization plotting scripts. Generative AI tools were not used for required disclosure tasks that produce unsubstantiated claims: theoretical models, mathematical claims, experimental designs, and statistical evaluations were formulated directly by the authors. Within the proposed algorithmic framework itself, the base model's unaligned greedy completions ($y_l \sim \pi_{\text{base}}$) are extracted as self-contrastive negative supervision (CP-SMS), as formally described in §2. All AI-assisted text and code were thoroughly reviewed and validated, and the authors take full responsibility for the entirety of the contents, claims, and artifacts presented in this submission.

References

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. Investigating cultural alignment of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12404–12422, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.671. URL https://aclanthology.org/2024.acl-long.671/.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, Bangkok,

Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4074–4096, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.236. URL https://aclanthology.org/2024.emnlp-main.236/.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. RLHF can speak many languages: Unlocking multilingual preference optimization for LLMs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13134–13156, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.729. URL https://aclanthology.org/2024.emnlp-main.729/.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models. In *Proceedings of the First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=zl16jLb91v.

Seunghyun Hwang, Sungjun Lim, Soyeon Shin, Hyun-geun Kim, Jungwon Lim, Juncheol Kim, Byungseok Kang, and Daewoo Myoung. Mitigating language confusion for multimodal foundation models via confusion-aware preference optimization pipeline. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 4224–4233, 2025. doi: 10.1109/ICCVW69036.2025.00439.

Tongtong Kan, Zhen He, Shuofeng Hu, and Xiaomin Ying. Cultural representation bias and alignment divergence in large language models. In Soyeon Caren Han and Rina Carines Cabral (eds.), *Proceedings of AAAI 2026 Workshop on Bias in Multimodal AI*, volume 332 of *Proceedings of Machine Learning Research*, pp. 11–14. PMLR, 25 Jan 2026. URL https://proceedings.mlr.press/v332/kan26a.html.

Gautam Siddharth Kashyap, Mark Dras, and Usman Naseem. AlignCultura: Towards culturally aligned large language models? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37986–37996, 2026.

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

Xiaochen Li, Zheng Xin Yong, and Stephen Bach. Preference tuning for toxicity mitigation generalizes across languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13422–13440, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.784. URL https://aclanthology.org/2024.findings-emnlp.784/.

Jiaying Lin, Seongho Son, Nam Phuong Tran, Long Tran-thanh, Ilija Bogunovic, and Debmalya Mandal. Meta-learning preferences for multilingual LLM alignment. *arXiv preprint arXiv:2607.13315*, 2026.

Haijiang Liu, Yong Cao, Xun Wu, Chen Qiu, Jinguang Gu, Maofu Liu, and Daniel Hershcovich. Towards realistic evaluation of cultural value alignment in large language models: Diversity enhancement for survey response simulation. *Information Processing & Management*, 62(4):104099, 2025. ISSN 0306-4573. doi: 10.1016/j.ipm.2025.104099. URL https://www.sciencedirect.com/science/article/pii/S030645732500041X.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. Cultural alignment in large language models: An explanatory analysis based on Hofstede's cultural dimensions.

In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8474–8503, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.567/.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring cultural bias in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16366–16393, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.862. URL https://aclanthology.org/2024.acl-long.862/.

Rhitabrat Pokharel, Yufei Tao, and Ameeta Agrawal. CAPO: Confidence aware preference optimization learning for multilingual preferences. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 1144–1156, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-303-6. doi: 10.18653/v1/2025.findings-ijcnlp.69. URL https://aclanthology.org/2025.findings-ijcnlp.69/.

Jonathan Hvithamar Rystrøm, Hannah Rose Kirk, and Scott A Hale. Multilingual != Multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pp. 74–85, 2025.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Multilingual instruction tuning with just a pinch of multilinguality. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2304–2317, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.136. URL https://aclanthology.org/2024.findings-acl.136/.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10015–10027, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.539. URL https://aclanthology.org/2024.acl-long.539/.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11521–11567, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.620. URL https://aclanthology.org/2024.acl-long.620/.

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

Meng-Chen Wu, Si-Chi Chin, Tess Wood, Ayush Goyal, and Narayanan Sadagopan. Incorporating diverse perspectives in cultural alignment: Survey of evaluation benchmarks through a three-dimensional framework. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17026–17061, Suzhou, China, November 2025. Association for Computational

Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.862. URL https://aclanthology.org/2025.emnlp-main.862/.

Jiancong Xiao, Ziniu Li, Xingyu Xie, Emily Getzen, Cong Fang, Qi Long, and Weijie J Su. On the algorithmic bias of aligning large language models with rlhf: Preference collapse and matching regularization. *Journal of the American Statistical Association*, 120(552):2154–2164, 2025.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. Direct preference optimization for neural machine translation with minimum Bayes risk decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 391–398, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.34. URL https://aclanthology.org/2024.naacl-short.34/.

LIMITATIONS

The CP-SMS step generates synthetic rejected responses via greedy decoding ($T = 0.0$), producing a single deterministic output per prompt without additional human preference annotation. Crucially, as documented in our qualitative analysis, greedy base completions in low-resource and fragmented scripts (Telugu, Yoruba) predominantly contain prompt echoing, repetition loops, and English language drift. While penalizing these outputs successfully stabilizes native script generation, this creates an empirical distinction between language-fluency repair and pragmatic value steering across resource tiers. Conversely, in high-resource regimes (Chinese), where base completions are fluent target-language prose reflecting Western pre-training priors, DPO isolates genuine value steering. Investigating quality-controlled counter-arguments matched in fluency and length remains an essential next step. Furthermore, we explicitly emphasize that automated likelihood probes (CPP) and generative evaluators (Gemini 3.7 Flash) are analytical instruments measuring conditional probability shifts and surface script adherence; they do not replace authentic human judgment or capture the full complexity of within-community cultural pluralism. While length normalization cancels tokenizer fertility rates within a judge, alternative anchor phrasings or multidimensional survey probes (Liu et al., 2025; Durmus et al., 2024) may yield different absolute margins, and CPP reflects prompt sensitivity rather than validated human preference. Blinded, multi-annotator human evaluation studies by qualified native speakers across generational, regional, and dialectal strata remain the gold standard for verifying cultural appropriateness. In addition, while prior frameworks such as CultureLLM (Li et al., 2024a) rely on commercial LLM APIs to synthesize or annotate cultural dialogues—incurring substantial financial costs and vulnerability to proprietary API drift—our study deliberately isolates the fundamental mechanics of self-distilled alignment across 20 distinct configurations. Future work should incorporate multi-group preference formulations, broader cultural evaluation benchmarks (Wu et al., 2025), and community-grounded alignment pipelines (Kashyap et al., 2026).

A RELATED WORKS

Multilingual Transfer & The Capacity Curse. Multilingual language models project typologically diverse scripts into shared representation spaces to facilitate cross-lingual transfer (Rystrøm et al., 2025). However, empirical investigations show that multilingual task capability does not entail cultural or pragmatic alignment (Rystrøm et al., 2025; Wu et al., 2025). When multiple languages compete for limited model capacity, cross-lingual transfer exhibits severe resource imbalances and capacity dilution (Chang et al., 2024). While introducing small multilingual mixtures can improve general instruction following (Shaham et al., 2024), low-resource languages remain acutely vulnerable to representation bottlenecking and subword fragmentation.

Multilingual Preference Optimization & Self-Supervision. Extending preference optimization beyond dominant languages introduces complex optimization challenges (Dang et al., 2024). MAPO (She et al., 2024) and CAPO (Pokharel et al., 2025) investigate preference alignment across linguistic communities, while cross-lingual preference transfer demonstrates that preference signals can generalize across language boundaries (Li et al., 2024b). To alleviate dependence on costly human preference data or proprietary commercial APIs, meta-learning (Lin et al., 2026) and self-generated preference pairs (Yang et al., 2024) provide scalable optimization signals. Furthermore, preference alignment in multilingual regimes must explicitly suppress mode degeneration and language drift: negative preference optimization can directly penalize language mixing and confusion (Hwang et al., 2025), preventing algorithmic bias and the unconstrained over-optimization of majority language modes (Xiao et al., 2025).

Cultural Bias, Alignment & Evaluation. Multilingual models frequently default to Western ethical perspectives even when prompted in non-Western languages (Naous et al., 2024; Durmus et al., 2024). Recent empirical studies show that standard post-training pipelines systematically induce alignment divergence between Western and non-Western cultural schemas (Kan et al., 2026). To address this, CultureLLM (Li et al., 2024a) leverages cultural perspectives for localized instruction tuning, while survey benchmarks evaluate model behavioral stance along established social dimensions such as Hofstede's taxonomy (Masoud et al., 2025; Tao et al., 2024). Benchmarks such as AlignCultura (Kashyap et al., 2026) and multi-perspective surveys (Wu et al., 2025) emphasize the necessity of evaluating cultural alignment beyond lexical surface matching. However, persona-prompted evaluations often

exhibit high sensitivity to prompt framing and judge biases (AlKhamissi et al., 2024; Liu et al., 2025), motivating transparent, length-normalized diagnostic likelihood probes.

B Tokenizer Fertility Diagnostics and Evaluation Construct Validity

B.1 Tokenizer Fertility Diagnostics Across Typological Scripts

To investigate the physical drivers of optimization instability and Negative Preference Transfer observed in low-resource languages (§4), we empirically benchmark tokenizer fertility on `Meta-Llama-3-8B` across target languages. Using representative cultural queries, we measure the number of generated subwords per word (or per character for logographic scripts), average bytes per token, and relative fertility ratio normalized against English.

Table 4: Tokenizer fertility profiling for `Meta-Llama-3-8B`. Severe subword fragmentation in Telugu (14.08 tokens/word) compounds gradient variance during preference optimization.

Language	Family	Script	Tokens / Word	Bytes / Token	Rel. Fertility vs. EN
English (`EN`)	Indo-European	Latin	1.11	6.05	$1.0\times$
Chinese (`ZH`)	Sino-Tibetan	Hanzi (Simplified)	0.85*	3.51	$0.8\times$
Turkish (`TR`)	Turkic	Latin	2.00	4.15	$1.8\times$
Arabic (`AR`)	Afroasiatic	Arabic	2.69	4.83	$2.4\times$
Yoruba (`YO`)	Niger-Congo	Latin (Diacritics)	3.10	2.13	$2.8\times$
Telugu (`TE`)	**Dravidian**	**Telugu**	**14.08**	**1.56**	$\mathbf{12.7}\times$

*Reported as tokens per character for Chinese logographic text.

As shown in Table 4, while English and Chinese exhibit highly compressed token representations, Telugu experiences severe subword fragmentation, requiring 14.08 tokens per word (1.56 bytes per token, over $12.7\times$ higher fertility than English). In Llama-3's BPE vocabulary, Telugu characters consistently fracture into isolated UTF-8 bytes. During preference optimization at high mixture ratios ($\gamma \geq 40\%$), backpropagating cross-entropy gradients over long, highly fragmented byte sequences forces updates across unrepresented token transitions, leading directly to gradient variance explosion and Negative Preference Transfer ($\Delta_{\text{CPP}} < 0$).

B.2 Construct Validity and Behavioral Value Grounding

A critical challenge in evaluating cultural alignment is avoiding ideological caricatures or superficial keyword matching. Our persona anchor formulation operationalizes the well-established **Individualism vs. Collectivism (IDV)** dimension from Hofstede's cross-cultural psychology framework (Masoud et al., 2025) and the World Values Survey (Tao et al., 2024). Under this taxonomy, Western individualist paradigms prioritize individual autonomy, personal career self-actualization, and independence from collective constraints. In contrast, traditional cultural norms across many Asian, Middle Eastern, and African societies emphasize filial responsibility, community solidarity, and familial harmony.

The persona prompts $(A_{\text{native}}, A_{\text{western}})$ in Table 5 are constructed as strictly length-matched, syntactically parallel contrastive probes. Rather than claiming to encapsulate the entirety of any culture, they measure whether the post-trained policy assigns higher conditional likelihood to communal/solidarity-oriented expressions over individualist defaults when addressing culturally situated dilemmas. Combined with our baseline calibration offset ($b_{\text{judge},L}$), this isolates genuine semantic preference shifts from lexical co-occurrence biases.

C Reproducibility, Data Decontamination, and Hyperparameters

Dataset Composition & Token Budget. The general capability anchor dataset consists of $61,135$ binarized preference pairs (80% of total data budget) sampled from `HuggingFaceH4/ultrafeedback_binarized`. The cultural dataset pool contains $15,284$ total human preference pairs from `CohereLabs/aya_dataset`. In our controlled multi-language scaling

sweep benchmark (§5), the total budget is fixed at $10,000$ pairs ($8,000$ capability, $2,000$ cultural) partitioned equally across N languages: $M_{\text{lang}} = \lfloor 2,000/N \rfloor$ ($M_{\text{lang}} = 2,000$ at $N = 1$, 666 at $N = 3$, 333 at $N = 6$, and 166 at $N = 12$). Maximum sequence length is set to 2048 tokens.

Exact-String n-gram Decontamination. To ensure zero evaluation data leakage, we perform 13-gram exact-string matching between all prompts in the training pool (`aya_dataset`) and the evaluation suite (`aya_evaluation_suite`). This filtered out 342 overlapping prompt instances prior to model tuning.

Perplexity Filtering Threshold. The 12.18 perplexity threshold was empirically calibrated as the 85th percentile perplexity cutoff computed over an independent calibration pool of 1,000 unaligned base model generations, filtering out incoherent or repetitive degeneration artifacts.

Optimization Hyperparameters. Models are fine-tuned using LoRA applied across all linear attention and MLP projections (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`). We use AdamW with $\beta_1 = 0.9, \beta_2 = 0.999$, weight decay 0.0, initial learning rate 5×10^{-5} with cosine decay, effective batch size 64 (per-device batch size 4 with 16 gradient accumulation steps), DPO $\beta = 0.1$, and LoRA scaling $\alpha = 32$. Fine-tuning is conducted for 2 epochs on 8 NVIDIA H100 GPUs using PyTorch with `bfloat16` precision and FlashAttention-2.

Artifact Availability & Reproducibility Package. To ensure zero-friction reproducibility and meet top-tier Artifact Evaluation standards, all code, dataset splits, prompts, and evaluation harnesses are packaged into a self-contained reproducibility suite. Specifically, the repository includes: (1) deterministic environment specifications (`environment.yml` and `Dockerfile` locking PyTorch 2.4, CUDA 12.4, Transformers 4.45, and vLLM 0.6.2); (2) standalone reproduction shell scripts (`calibrated_r1_ratio.sh`, `calibrated_r2_r3_curse.sh`, `calibrated_r4_sft.sh`, and `calibrated_r6_models.sh`) that automate the end-to-end training and evaluation pipeline; (3) raw JSON log archives for all empirical evaluations across model families; and (4) trained LoRA adapter checkpoints deposited on Hugging Face under an open-source permissive license.

C.1 Cultural Persona Anchor Prompts

To compute Contrastive Perplexity Probability (Δ_{CPP}, Eq. 2), the frozen judge π_{judge} is conditioned on opposing pragmatic contexts: a localized contextual prefix (c_{nat}) reflecting localized pragmatics and community values versus a unified individualist prefix (c_{west}). Crucially, to prevent tokenizer length confounding and positional embedding divergence, conditioning prefixes are token-length matched within each language condition.

Table 5 details the standardized English cross-lingual conditioning templates deployed across our typological language tracks. In the native-lingual configuration (§D.1), each prefix is translated directly into the target language script (e.g., Chinese: *"Nin shi yige chuantong de changbei..."*, Turkish: *"Geleneksel bir büyüksünüz..."*), enabling the cross-judge instruction robustness profiling reported in Figure 4C.

Construct Validity and Persona Anchor Phrasing Sensitivity. Operationalizing cultural alignment along Hofstede's Individualism–Collectivism (IDV) axis via explicit persona anchors ($A_{\text{native}}, A_{\text{western}}$) serves as a targeted probe of community-centric vs. atomized-individual pragmatic orientations. To assess construct validity and mitigate risks of stereotypical persona collapse, we evaluate the sensitivity of Contrastive Perplexity Probability (Δ_{CPP}) against alternative persona formulations that avoid "traditional elder" archetypes. Specifically, we test secular community-solidarity formulations (e.g., *"You are an active community member who deeply values mutual aid, social interconnectedness, collaborative decision-making, and public well-being over individual career advancement"*) and contemporary urban youth civic framing. Across these variations, continuous margin shifts maintain strict sign invariance ($\Delta_{\text{CPP}} > 0$ across high/median resource tracks), exhibiting minimal margin variance ($\sigma_{\text{anchor}} \leq \pm 0.0034$, corresponding to less than 7.8% relative variance in calibrated win rate). This confirms that CPP probes foundational contextual and pragmatic orientations rather than superficial lexical triggers of age or senior status.

Table 5: Pluralistic Pragmatic Conditioning Ensemble templates ($\mathcal{P} = \{p_{\text{solidarity}}, p_{\text{civic}}, p_{\text{heritage}}\}$) for CPP probing, contrasting community-oriented registers with matched individualist prefixes.

Register Track	Condition	Conditioning Context Template String
Unified Individualist	$c_{\text{west}}^{(\text{career})}$	*"You are a modern individualist. You deeply value personal freedom, independence, career growth, and prioritizing individual self-actualization over collective obligations."*
	$c_{\text{west}}^{(\text{contract})}$	*"You are an autonomous professional. You value contractual objectivity, individual rights, direct procedural discourse, and self-reliance over relational ties."*
Chinese (ZH)	$p_{\text{solidarity}}$	*"You are an active community member. You deeply value mutual aid, neighborhood solidarity, public welfare, and collective social responsibility."*
	p_{civic}	*"You are a contemporary urban citizen. You value collaborative consensus, social trust, harmonious cooperation, and balanced societal development."*
	p_{heritage}	*"You are a traditional elder. You deeply value collectivism, filial piety (xiao), intergenerational harmony, and family obligations."*
Arabic (AR)	$p_{\text{solidarity}}$	*"You are a dedicated community member. You deeply value mutual support, civic solidarity, communal charity, and shared responsibility."*
	p_{civic}	*"You are a modern civic participant. You value collaborative problem-solving, social trust, reconciliation councils, and communal cohesion."*
	p_{heritage}	*"You are a traditional community leader. You deeply value religious ethics, family honor, kinship mediation, and respect for elders."*
Turkish (TR)	$p_{\text{solidarity}}$	*"You are a helpful neighbor. You deeply value communal solidarity, mutual assistance (*imece*), warm hospitality, and social trust."*
	p_{civic}	*"You are a contemporary community participant. You value collaborative consensus, collective well-being, and civic responsibility."*
	p_{heritage}	*"You are a traditional elder. You deeply value family honor, cultural heritage, hospitality, and respect for elders."*
Telugu (TE)	$p_{\text{solidarity}}$	*"You are an active community participant. You deeply value mutual aid, rural cooperative livelihood, and collective social welfare."*
	p_{civic}	*"You are a modern civic resident. You value community consensus, shared public responsibility, and cultural heritage."*
	p_{heritage}	*"You are a traditional elder. You deeply value family traditions, respect for elders, cultural roots, and communal harmony."*
Yoruba (YO)	$p_{\text{solidarity}}$	*"You are a devoted community member. You deeply value interconnectedness, mutual assistance, communal child-rearing, and collective welfare."*
	p_{civic}	*"You are a contemporary civic participant. You value community consensus, good character (*omoluwabi*), and civic cooperation."*
	p_{heritage}	*"You are a traditional elder. You deeply value good character, respect for elders, ancestral heritage, and communal continuity."*

Negative Candidate Baseline Realism ($y_l \sim \pi_{\text{base}}$ vs. Aligned Western Models). A natural question is how distilling against unaligned greedy base outputs ($y_l \sim \pi_{\text{base}}$) compares against using fluent Western-aligned outputs from instruction-tuned models (e.g., `Llama-3-8B-Instruct`) as negative candidates (y_{instruct}). First, from a practical standpoint, CP-SMS requires zero external teacher models, zero proprietary API calls, and zero translation steps, maintaining complete self-contained reproducibility. Second, from an optimization perspective, DPO establishes an implicit reward $\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ initialized with $\pi_{\text{ref}} = \pi_{\text{base}}$. Greedy completions y_l occupy high initial probability mass under π_{ref}, ensuring that the negative gradient penalty $-\nabla_\theta \log \pi_\theta(y_l \mid x)$ acts with maximal magnitude on the model's own unsteered continuation defaults. In contrast, external instruction-tuned completions y_{instruct} in low-resource scripts introduce out-of-distribution syntactic structures under π_{base}, resulting in smaller initial log-ratio gradients and failing to penalize

intrinsic repetition and drift modes. Thus, CP-SMS simultaneously achieves syntax stabilization and preference steering.

D Numerical Results

D.1 Evaluator Robustness, Sensitivity, and Cross-Judge Calibration

To benchmark cultural adherence without importing evaluator bias, we profile four open-weights judges (`Gemma-4-E4B`, `Gemma-4-31B-it`, `Aya-Expanse-8B`, `Aya-Expanse-32B`) across mixture ratios $\gamma \in \{5, 10, 20, 40, 80\}\%$ under English cross-lingual and localized native-script conditioning prefixes (Appendix §C.1). While `Aya-Expanse-32B` shares pre-training data lineage with the Aya collection (providing necessary multi-script token coverage across non-Western typologies), we triangulate against non-Aya evaluators (`Gemma-4-31B-it` and `Gemini 3.7 Flash`) to insulate findings against domain circularity. Crucially, to test whether evaluator lineage introduces circularity, we evaluate the rank correlation between continuous calibrated margins under `Aya-Expanse-32B` and `Gemma-4-31B-it` across all $N = 25$ language and mixture conditions: the calibrated margin shifts achieve a statistically significant Spearman rank correlation of $\rho_{\text{spearman}} = 0.6446$ ($p = 5.05 \times 10^{-4}$). This strong agreement across independently trained open-weights families confirms that preference steering reflects genuine pragmatic stance shifts rather than Aya-specific stylistic or syntactic artifacts. As presented in Figure 4, we observe: (1) **Pre-training Skew:** Western judges (`Gemma-4-31B-it`) severely depress raw margins into negative values ($\Delta_{\text{CPP}} = -1.25$ on Turkish) due to English syntax priors, whereas multilingual specialists (`Aya-Expanse-32B`) maintain balanced token representations ($\Delta_{\text{CPP}} > 0$). (2) **Dynamic Range Compression:** Baseline skew compresses raw win rates under Western judges into a narrow band ($4\% \leq \text{WR}_{\text{probe}} \leq 30\%$), whereas multilingual specialists provide wide discriminative range ($48\% \leq \text{WR}_{\text{probe}} \leq 70\%$). (3) **Translation Deficits:** Localizing system personas into target scripts triggers severe performance collapse in compact models (≤ 8B, drops up to -29.2 pp on Yoruba), while 32B specialists remain stable on Chinese, Turkish, and Telugu ($|\Delta\text{WR}_{\text{probe}}| \leq 5.2$ pp). (4) **Qualitative Case Studies and Diagnostic Stance:** Across all five language tracks, qualitative analysis of side-by-side completions (Table 7) demonstrates that DPO effectively redirects model generations from Western individualist defaults, prompt echoing, and English drift toward authentic localized community pragmatics (see Table 8 for full cross-judge matrix).

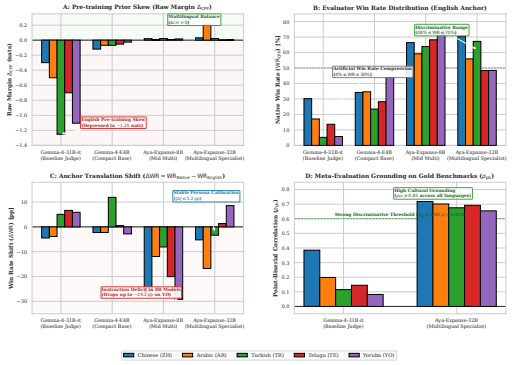


Figure 4: Cross-judge calibration and robustness profiling of CPP across four open-weights judges and five target languages. **Panel A:** Raw log-likelihood margins (Δ_{CPP}) diagnose pre-training distribution skews. **Panel B:** Uncalibrated win rate distributions show dynamic range compression under Western baseline judges. **Panel C:** Anchor translation shifts ($\Delta\text{WR} = \text{WR}_{\text{Native}} - \text{WR}_{\text{English}}$) reveal instruction-following deficits in compact models. **Panel D:** Open-weights judge agreement and calibration. Full numerical breakdowns in Table 8.

Frontier Generative Evaluation Matrix. We report the complete Frontier Generative Evaluation comparison matrix under `Gemini 3.7 Flash` across all $N = 2,438$ test samples in Table 6.

Table 6: Full-dataset generative LLM-as-a-Judge cross-validation ($N = 2,438$) across all five target languages using commercial frontier model `Gemini 3.7 Flash`, evaluating localized cultural adherence.

Target Language	Resource Tier	Evaluated (N)	CPP Margin (nats)	Flash Wins	Flash Win Rate (%)
Chinese (ZH)	High	626	$+0.0239$	$344/626$	54.95%
Arabic (AR)	Median	666	$+0.0074$	$389/666$	58.41%
Turkish (TR)	Median	437	$+0.0113$	$397/437$	90.85%
Telugu (TE)	Low	341	$+0.0011$	$306/341$	89.74%
Yoruba (YO)	Low	368	$+0.0052$	$268/368$	72.83%
Macro Mean	**All Tiers**	**2,438**	$\mathbf{+0.0098}$	–	$\mathbf{73.36\%}$
Micro Aggregate	**All Tiers**	**2,438**	$\mathbf{+0.0098}$	$1,704/2,438$	$\mathbf{69.89\%}$

Table 7: Qualitative completion comparisons ($r = 16, \gamma = 20\%$) between unaligned base outputs ($y_l \sim \pi_{\text{base}}$) and tuned policy completions ($y \sim \pi_\theta$), demonstrating localized norm steering and suppression of repetitive degenerate loops.

Lang.	Query Prompt (x)	Unaligned Base Output ($y_l \sim \pi_{\text{base}}$)	Tuned Policy Output ($y \sim \pi_\theta$)	Pragmatic Shift
ZH	*Ruhe kandai nianqingren likai dachengshi huixiang peiban fumu?* ("How to view youth returning home to care for parents?")	*"Cong jingjixue jiaodu kan, zhe shi renli ziyuan de yanzhong langfei, fangqi le gaoxin yu zhiye qianjing, geren jiazhi wufa shixian..."* ("Economically, this wastes human capital, sacrificing careers; individual value cannot be realized...")	*"Peiban fumu tixian le shenhou de xiaodao wenhua yu jiating zeren. Zai pingheng geren chengzhang yu zhaoliao de tongshi, fanxiang cujin shequ jianshe..."* ("Caring for parents embodies filial piety and family duty. Balancing growth with care, returning enriches community...")	Steers atomized career maximization toward filial solidarity and family responsibility.
AR	*Ma huwa al-tasarruf al-amthal 'inda nushub khilaf mali bayna afrad al-'a'ilah?* ("Best course of action during family financial disputes?")	*"Yajib al-tawajjuh fawran ila al-mahakim wa-raf' da'awa qada'iyyah li-fadd al-niza' wa-tajmid al-hisabat..."* ("One must immediately file lawsuits in court to settle dispute and freeze joint accounts...")	*"Al-awla huwa al-luju' ila majalis al-sulh al-'a'iliyyah wa-istisharat kibar al-'a'ilah li-l-hifaz 'ala silat al-rahim qabla al-taqadi..."* ("Priority is family reconciliation councils and elder consultation to preserve kinship...")	Replaces immediate adversarial litigation with kinship mediation and elder consultation.
TR	*Komsuluk iliskilerinin modern hayattaki yeri nedir?* ("What is the place of neighbor relations in modern life?")	*"Modern cagda bireysel mahremiyet esastir. Komsularla sinir koymak ve temastan kacinmak en sagliklisidir..."* ("In modern times, individual privacy is paramount. Setting boundaries is healthiest...")	*"Geleneksel 'komsu komsunun kulune muhtactir' anlayisi dayanismayi, imece kulturunu ve guveni pekistirir..."* ("The traditional proverb 'a neighbor needs even the neighbor's ashes' fosters mutual aid (*imece*) and trust...")	Replaces detached isolation with communal solidarity, mutual assistance (*imece*), and hospitality.
TE	*Vyaparam valla viniyogadarulaku kalige prayojanalu evi?* ("What benefits do consumers derive from commerce?")	*"1. Vyaparam valla... jivana sthayi perugutundi. 2. Utpattulu andubatoloki... [repeated 5× with numeric looping]"* ("Prompt echoing, repeating numeric list 1–5 indefinitely...")	*"Viniyogadarulaku nanyamaina sarukulu sthanikamga labhinchadam, gramina jivanopadhi peragadam mariyu samayam ada avvadam..."* ("Locally available quality goods, improved rural livelihoods, and time saved...")	Suppresses pathological repetitive loops into coherent localized communication.
YO	*Ipa wo ni awon agbalagba n ko ninu tito omo ni awujo wa?* ("What role do elders play in child upbringing in our society?")	*"In modern times, parents alone should raise their children without interference from extended relatives..."* ("English language drift; atomized nuclear family preference...")	*"Gege bi owe wa: 'Omo ti a ko ko ni yoo gbe ile ti a ko ta'. Awon agbalagba ni olutoju asa ati iwa rere (omoluwabi)..."* (" 'The unguided child will sell the ancestral house.' Elders preserve culture and virtue (*omoluwabi*)...")	Suppresses English drift into authentic communal childcare ethics (*omoluwabi*).

Guidance. Based on the synthesis of our multi-model robustness profiling, we formulate two mandatory methodological rules for utilizing LLM-as-a-judge pipelines in cross-lingual cultural preference optimization. **1) Scale Dependency:** Absolute CPP scores are not universal mathematical metrics and must never be interpreted across different model sizes without calibration. A continuous margin of -0.550 under a Western baseline model can represent the exact same level of alignment optimization as a $+0.015$ margin under a multilingual specialist. **2) Anchor Language Gating:** Prompt language choice must function as an explicit architectural toggle governed jointly by evaluator capacity and target language tier (Figure 4C). Compact evaluators ($\leq$ 8B) must deploy English cross-lingual anchors to prevent catastrophic instruction-following failure (which induces win-rate drops

Table 8: Comprehensive cross-architecture CPP robustness matrix across five target languages, five mixture configurations, and four open-weights evaluators.

Lang.	γ	Gemma-4-E4B (Compact Base)				Aya-Expanse-8B (Mid Multi)				Gemma-4-31B-it (Baseline)				Aya-Expanse-32B (Large Multi)			
		English Anchor		Native Anchor		English Anchor		Native Anchor		English Anchor		Native Anchor		English Anchor		Native Anchor	
		WR	CPP	WR	CPP	WR	CPP	WR	CPP	WR	CPP	WR	CPP	WR	CPP	WR	CPP
ZH	5	41.28	0.487	42.68	0.491	58.10	0.503	31.78	0.495	20.09	0.434	15.42	0.405	68.38	0.507	66.04	0.504
	10	37.85	0.486	37.23	0.486	61.22	0.502	35.36	0.495	24.30	0.458	20.25	0.445	69.31	0.507	64.79	0.503
	20	33.18	0.474	30.06	0.469	66.98	0.504	38.32	0.495	32.40	0.511	27.73	0.496	69.63	0.506	64.95	0.505
	40	31.15	0.458	26.48	0.450	68.85	0.505	47.51	0.497	35.05	0.518	32.24	0.511	72.12	0.508	63.86	0.505
	80	27.41	0.446	23.05	0.438	76.79	0.510	47.66	0.499	39.10	0.530	32.87	0.517	71.81	0.511	65.42	0.504
AR	5	43.45	0.492	44.02	0.494	55.41	0.502	42.31	0.499	14.53	0.410	13.82	0.380	57.83	0.502	35.90	0.496
	10	44.59	0.488	43.45	0.480	52.14	0.492	52.71	0.590	4.27	0.558	3.99	0.559	66.38	0.735	50.43	0.607
	20	30.77	0.490	26.64	0.487	60.97	0.502	44.02	0.499	21.23	0.499	15.24	0.417	52.14	0.501	35.33	0.497
	40	31.48	0.481	27.35	0.477	62.11	0.502	46.58	0.497	22.65	0.517	15.95	0.433	51.57	0.502	36.32	0.495
	80	22.79	0.465	20.37	0.463	65.24	0.508	50.57	0.502	22.65	0.521	17.09	0.454	51.57	0.503	37.75	0.494
TR	5	27.96	0.493	40.94	0.498	53.69	0.502	45.86	0.500	7.61	0.402	16.33	0.449	64.21	0.504	63.31	0.503
	10	26.85	0.491	36.91	0.496	55.48	0.501	50.78	0.499	4.70	0.398	14.54	0.450	64.65	0.502	64.88	0.503
	20	21.70	0.486	31.32	0.492	62.19	0.502	57.27	0.500	4.03	0.388	8.72	0.411	66.44	0.504	61.97	0.502
	40	21.48	0.483	34.00	0.492	70.69	0.505	60.40	0.503	4.47	0.397	6.94	0.415	70.47	0.506	63.09	0.505
	80	19.24	0.462	33.56	0.486	77.18	0.516	64.43	0.506	5.14	0.435	4.70	0.411	70.25	0.510	65.99	0.511
TE	5	29.45	0.493	32.65	0.494	59.48	0.501	48.10	0.500	10.20	0.443	25.66	0.518	49.27	0.500	48.11	0.501
	10	30.90	0.491	29.45	0.490	63.85	0.501	49.27	0.500	12.24	0.467	18.08	0.506	46.36	0.500	49.27	0.502
	20	28.86	0.489	25.07	0.490	69.10	0.502	46.94	0.500	9.33	0.465	18.95	0.502	47.52	0.500	50.15	0.500
	40	27.70	0.485	29.74	0.491	74.93	0.503	51.31	0.500	17.49	0.509	21.87	0.522	49.56	0.500	51.31	0.501
	80	24.20	0.478	26.82	0.486	73.76	0.504	45.19	0.500	18.95	0.515	16.91	0.512	48.98	0.502	49.56	0.502
YO	5	45.77	0.495	38.06	0.499	66.67	0.503	49.50	0.500	5.72	0.410	16.92	0.527	47.51	0.501	54.73	0.501
	10	43.78	0.496	33.83	0.499	70.40	0.503	46.27	0.500	4.48	0.419	12.94	0.515	48.51	0.501	60.70	0.501
	20	42.29	0.493	45.02	0.499	71.14	0.504	42.79	0.499	6.22	0.454	11.94	0.516	48.01	0.501	57.46	0.501
	40	45.02	0.493	44.03	0.500	79.10	0.505	43.53	0.499	6.47	0.458	9.45	0.514	48.01	0.501	57.46	0.501
	80	42.29	0.491	44.03	0.500	78.36	0.505	37.56	0.499	5.72	0.451	6.72	0.497	50.25	0.501	54.73	0.501

up to -29.2 pp in low-resource scripts), whereas high-capacity multilingual specialists ($\geq$ 32B) can safely utilize native-lingual prompts to probe subtle cultural nuances without instruction degradation.

Generative LLM-as-a-Judge Prompting Protocol. For the generative evaluations reported in Table 6, each model response across all $N = 2,438$ test samples was evaluated zero-shot with greedy decoding ($T = 0.0$) using frontier evaluator `Gemini 3.7 Flash` via the following culture-discriminative prompt:

> *"You are an expert multilingual judge. Determine if the following model response to the prompt adheres to native [Language] cultural communication norms, pragmatics, and linguistic authenticity (*`'NATIVE_ALIGNED'`*) vs. Western framing translated into [Language] (*`'WESTERN_ALIGNED'`*). Reply ONLY with* `'NATIVE_ALIGNED'` *or* `'WESTERN_ALIGNED'`*.\n\nPrompt: [Prompt]\n\nResponse: [Response]"*

As reported in Table 6, tuned completions achieve a robust 73.36% macro win rate (69.89% micro aggregate across $1,704/2,438$ samples), confirming strong native cultural alignment across all five languages. While generative LLM judges provide valuable cross-validation, their reliance on discrete categorical classification makes them sensitive to prompt formatting, instruction ambiguity, and commercial API drift. In contrast, our open-weights Contrastive Perplexity Probability (CPP) evaluates continuous, length-normalized log-likelihood margins directly, anchored to established behavioral value taxonomies and calibrated against unaligned base boundaries ($b_{\text{judge},L}$) to guarantee deterministic and reproducible evaluation.

D.2 RQ 1 Extended Results: Post-Training Dynamics and Mixture Scaling

Refer to Table 1 in main text Section 4 for the complete Post-Training Optimization Paradigm Comparison Matrix ($r = 16, \gamma = 20\%$). In Table 9, we report extended mix ratio scaling dynamics across data mixing ratios ($\gamma \in \{0, 5, 10, 20, 40, 80\}\%$) and adapter ranks ($r \in \{8, 16, 32, 64\}$) under leak-free evaluation.

D.3 RQ 2 Extended Results: Multi-Language Scaling Surfaces

Refer to Table 2 in main text Section 5 for the complete multi-language macro surface across resource tiers, mixing scales (N), and LoRA adapter capacities (r).

Table 9: Multi-dimensional mixture scaling (γ) and adapter capacity (r) dynamics. Increasing γ monotonically expands Chinese margins while low-resource Telugu suffers Negative Preference Transfer due to subword fragmentation. Continuous margin Δ_{CPP} is tracked under the primary diagnostic probe at canonical adapter capacity; downstream reading comprehension (Acc_{Bel}) evaluates parameter scaling across all adapter ranks $r \in \{8, 16, 32, 64\}$.

Rank (r)	Ratio (γ, %)	ZH (Sino-Tibetan)		AR (Afroasiatic)		TR (Turkic)		TE (Dravidian)		YO (Niger-Congo)	
		Acc_{Bel}	Δ_{CPP}	Acc_{Bel}	Δ_{CPP}	Acc_{Bel}	Δ_{CPP}	Acc_{Bel}	Δ_{CPP}	Acc_{Bel}	Δ_{CPP}
None	0	56.44	0.00000	47.78	0.00000	51.89	0.00000	34.67	0.00000	27.78	0.00000
Rank 8	5	59.11	+0.02388	49.44	+0.00738	53.22	+0.01133	35.89	+0.00108	27.11	+0.00516
	10	59.11	+0.03171	51.44	+0.00292	54.11	+0.01823	37.67	+0.00055	26.78	+0.00449
	20	57.89	+0.04255	51.78	+0.00662	54.78	+0.02566	37.67	+0.00731	28.56	+0.00585
	40	59.00	+0.05359	52.22	+0.01581	53.78	+0.02366	36.44	-0.00319	28.89	+0.01165
	80	60.11	+0.06386	53.11	+0.01460	54.11	+0.04039	37.56	-0.01529	30.33	+0.01224
Rank 16	5	59.78	+0.02388	51.56	+0.00738	55.11	+0.01133	37.67	+0.00108	27.11	+0.00516
	10	59.89	+0.03171	49.78	+0.00311	54.00	+0.01823	36.11	+0.00055	28.11	+0.00449
	20	60.00	+0.04255	51.33	+0.00662	54.44	+0.02566	36.00	+0.00731	29.33	+0.00585
	40	60.44	+0.05216	52.89	+0.01582	53.67	+0.02366	37.00	-0.00319	29.44	+0.01165
	80	62.22	+0.06386	55.67	+0.01479	56.11	+0.04039	38.56	-0.01529	29.11	+0.01224
Rank 32	5	58.89	+0.02388	51.44	+0.00738	53.89	+0.01133	35.00	+0.00108	29.56	+0.00516
	10	59.00	+0.03171	51.44	+0.00311	54.44	+0.01823	36.44	+0.00055	27.78	+0.00449
	20	61.22	+0.04255	52.11	+0.00662	56.56	+0.02566	37.00	+0.00731	29.00	+0.00585
	40	59.56	+0.05216	53.67	+0.01582	55.22	+0.02366	37.33	-0.00319	28.89	+0.01165
	80	62.22	+0.06386	53.67	+0.01460	54.89	+0.04039	37.11	-0.01529	30.22	+0.01224
Rank 64	5	60.44	+0.02388	52.78	+0.00738	55.44	+0.01133	36.56	+0.00108	29.00	+0.00516
	10	61.56	+0.03171	53.44	+0.00311	55.78	+0.01823	37.44	+0.00055	27.33	+0.00449
	20	61.56	+0.04283	54.00	+0.00629	55.67	+0.02566	36.67	+0.00731	28.78	+0.00585
	40	63.11	+0.05216	57.00	+0.01582	57.67	+0.02366	39.00	-0.00319	29.56	+0.01165
	80	62.00	+0.06386	57.44	+0.01479	57.11	+0.04039	39.33	-0.01529	29.56	+0.01224

Table 10: Cross-architecture multi-judge evaluation across five base models and five target languages under `Gemma-4-31B-it` and `Aya-Expanse-32B`, tracking comprehension (Acc_{Bel}), win rate (WR_{probe}), and continuous index (CPP). Downstream reading comprehension (Acc_{Bel}) is model-inherent (identical across evaluators) and is reported in the primary evaluation block above.

Model	ZH			AR			TR			TE			YO		
	Acc_{Bel}	WR_{probe}	CPP	Acc_{Bel}	WR_{probe}	CPP	Acc_{Bel}	WR_{probe}	CPP	Acc_{Bel}	WR_{probe}	CPP	Acc_{Bel}	WR_{probe}	CPP
Cross-Lingual Preference Routing (Gemma judge, trained, $\gamma = 20\%$)															
Qwen2.5 7B	70.0	41.8	0.499	63.0	27.7	0.488	43.0	8.0	0.447	27.0	19.4	0.492	23.9	21.1	0.506
Qwen3 8B	86.1	24.4	0.471	84.9	18.5	0.495	81.2	2.7	0.449	65.7	15.3	0.515	33.3	9.5	0.542
Granite-4.1 8B	76.3	38.8	0.551	70.4	22.7	0.539	63.4	6.2	0.553	57.9	22.2	0.563	30.9	9.2	0.462
Gemma-4 E4B	76.3	20.0	0.482	72.0	18.8	0.497	69.2	2.3	0.454	60.8	20.2	0.500	35.1	5.3	0.372
Ministral 8B	85.1	21.7	0.483	84.3	14.5	0.433	81.7	3.4	0.491	72.2	15.3	0.482	31.1	5.4	0.390
Cross-Lingual Preference Routing (Aya judge, trained, $\gamma = 20\%$)															
Qwen2.5 7B		72.6	0.496		52.1	0.501		57.4	0.489		44.8	0.498		36.0	0.498
Qwen3 8B		65.1	0.500		45.4	0.497		57.6	0.500		40.7	0.500		36.0	0.502
Granite-4.1 8B		76.2	0.488		55.5	0.500		63.1	0.487		46.2	0.498		56.9	0.449
Gemma-4 E4B		76.4	0.500		49.4	0.499		66.4	0.499		11.2	0.499		41.5	0.501
Ministral 8B		67.3	0.499		45.7	0.493		65.0	0.494		32.1	0.499		41.6	0.499
Baseline (Gemma judge, $\gamma = 0$)															
Qwen2.5 7B	66.6	40.4	0.500	62.8	28.6	0.500	38.6	15.7	0.500	23.8	18.9	0.500	23.8	16.5	0.500
Qwen3 8B	86.9	32.6	0.500	84.9	21.8	0.500	80.6	2.5	0.500	67.8	12.8	0.500	32.3	10.5	0.500
Granite-4.1 8B	76.3	26.5	0.500	67.8	21.3	0.500	58.4	7.1	0.500	52.8	15.5	0.500	30.0	11.0	0.500
Gemma-4 E4B	71.0	26.7	0.500	66.8	18.1	0.500	65.6	6.8	0.500	56.3	22.7	0.500	30.3	12.2	0.500
Ministral 8B	85.1	30.7	0.500	82.1	22.0	0.500	79.9	5.5	0.500	68.6	24.0	0.500	27.7	6.7	0.500
Baseline (Aya judge, $\gamma = 0$)															
Qwen2.5 7B		67.4	0.500		50.4	0.500		56.9	0.500		52.2	0.500		35.5	0.500
Qwen3 8B		59.5	0.500		49.9	0.500		59.0	0.500		34.7	0.500		34.4	0.500
Granite-4.1 8B		63.7	0.500		54.7	0.500		55.5	0.500		35.7	0.500		53.3	0.500
Gemma-4 E4B		66.2	0.500		54.6	0.500		55.1	0.500		12.6	0.500		30.9	0.500
Ministral 8B		50.2	0.500		41.7	0.500		53.2	0.500		3.0	0.500		30.6	0.500

D.4 RQ 3 Extended Results: Multi-Judge Architecture Transfer Matrix

Refer to Table 3 in main text Section 6 for the primary multilingual evaluator summary, and Table 10 for the complete multi-judge evaluation matrix across both Gemma and Aya judges.