THE MULTILINGUAL ALIGNMENT PARADOX: PREFERENCE CONFLICT HURTS ALIGNMENT BUT PREVENTS LANGUAGE COLLAPSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Preserving language-specific preferences is challenging under language-imbalanced training, especially when preferences conflict across languages. We study this interaction using a controlled preference dataset derived from global opinion data, varying language imbalance and cross-language preference conflict. Across four multilingual LLMs, we compare supervised fine-tuning (SFT) and direct preference optimization (DPO). As imbalance increases, SFT largely preserves the input language but loses preference accuracy, whereas DPO increasingly produces English responses to minority-language inputs, especially under shared preferences. This reveals a *multilingual alignment paradox*: preference conflict reduces preference accuracy but can protect language fidelity. At an English-to-each-minority ratio of 100:1, conflict reduces English fallback across all four models and raises mean joint language-and-preference accuracy after DPO from 43.2% to 62.1%. Objective analysis shows that, unlike SFT, within-language DPO comparisons leave output-language probabilities unconstrained at the distribution level, while training dynamics indicate stronger residual minority preference signals under conflict. Sparse autoencoder analysis further identifies stronger target-language and weaker English-associated activity under conflict. Intervening on these features reduces English fallback with limited impact on preference accuracy. Together, these findings suggest that language-selection features retained with language-specific preferences even under severe imbalance in the alignment process.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed to serve users from diverse linguistic and cultural backgrounds (Qin et al., 2025). Despite this global reach, current LLMs have been shown disproportionately reflect Western-centric norms and values (Durmus et al., 2023; Ryan et al., 2024), because the underlying training data landscape remains highly imbalanced for different languages (W3Techs, 2026). This creates a fundamental *trade-off* in the multilingual preference alignment: leveraging high-resource languages improves shared knowledge transfer, but aslo risks suppressing language-specific preferences (Han et al., 2025). This issue becomes particularly severe under highly imbalanced settings, where minority-language preferences are easily overridden by majority-language preferences during optimization (Yao et al., 2025).

Prior multilingual alignment has mostly focused on expanding multilingual instruction-tuning resources (Muennighoff et al., 2022; Üstün et al., 2024),and transfer alignment signals from high-resource to lower-resource languages (She et al., 2024; Zhao et al., 2025; Yang et al., 2025b;c). Complementary work identifies shared semantic representations in multilingual LLMs (Wu et al., 2024) and shows that deeper-layer representations can support cultural localization and recover language-specific information through steering (Han et al., 2025). However, it's less known about how imbalanced multilingual preference data including both the shared and conflicting preferences affect the preference alignment process. We therefore ask:

How do preference conflicts under data imbalance affect multilingual preference alignment?

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.

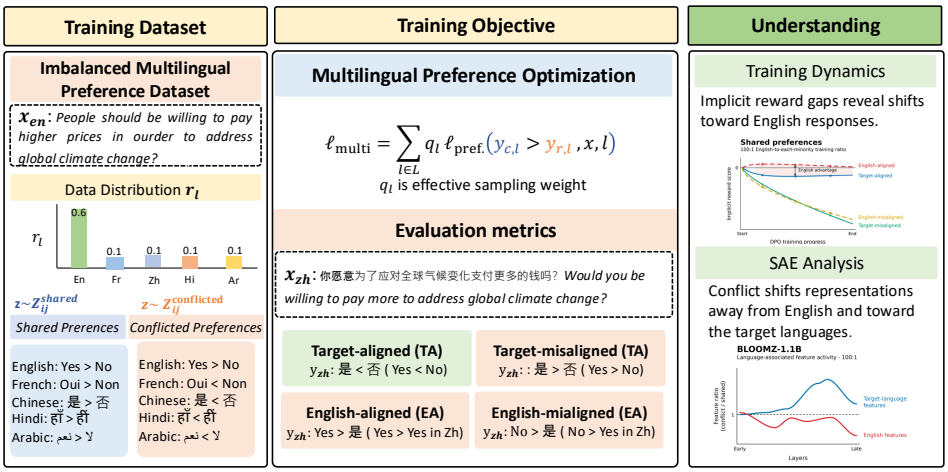


Figure 1: Overview of the imbalanced multilingual training dataset (`left`), training objective (`middle`), and insights from training dynamics and SAE representations (`right`).

To isolate the factors of data imbalance and preference conflict, we construct controlled synthetic preference datasets derived from GlobalOpinionQA, a dataset about subjective opinions on global issues (Durmus et al., 2023), covering English, Chinese, French, Arabic, and Hindi (shown in Figure 1). To control data imbalance, we designate English as the dominant language and vary the English-to-each-minority training-data ratio. For preference conflict, we construct two partitions: To control preference conflict, we construct two partitions: a shared-preference partition, in which English preferences are translated into each minority language while preserving the preference ordering, and a conflicting-preference partition, in which all minority-language preferences reverse the English ordering. Across four multilingual LLMs, we evaluate supervised fine-tuning (SFT) and subsequent direct preference optimization (DPO) to compare learning behavior across training stages.

Our experiments reveal a ***multilingual alignment paradox***: *preference conflict hurts alignment but prevents language collapse.* As imbalance increases, SFT largely preserves the input language but makes more preference errors. DPO introduces substantial English fallback when preferences are shared. However, conflicting preferences, although make preference selection harder, reduce English fallback across all four models under imbalanced training data. Averaged across models and minority languages, conflict lowers joint language-and-preference accuracy after SFT from 77.0% to 70.0%, but raises it after DPO from 43.2% to 62.1%. These results expose a tension between satisfying preferences and preserving the language used to express them.

We investigate this tension through objective analysis, training dynamics, and representation interventions. In within-language DPO comparisons, the probability factors associated with the response language cancel, whereas SFT includes an explicit target-language likelihood term. A local surrogate illustrates how conflict can sustain minority-conditioned preference corrections, consistent with the larger residual minority DPO weights observed under severe imbalance. Sparse autoencoder analysis further associates conflict with stronger target-language-associated activity and weaker English-associated activity. Amplifying target-language-associated features reduces English fallback to near zero for most models in both shared- and conflicting-preference settings under severe imbalance, with limited cost to preference correctness. These interventions provide causal evidence that the identified representations influence output-language behavior. Together, the findings suggest that language-selection features retained with language-specific preferences, even under severe imbalance.

2 Related Work

Preference Pluralism in Multilingual Alignment. Most preference-alignment methods, including RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023), optimize preference signals aggregated across users or examples. This assumption becomes limiting in multilingual and multicultural settings, where preferences may legitimately differ across linguistic and cultural communities (Han et al., 2025; Yao et al., 2025). Prior work documents substantial cross-cultural variation in normative judgments (Awad et al., 2018; Takemoto, 2024) and argues that language technologies often reflect dominant populations and norms (Bender et al., 2021). These findings motivate preserving cross-language disagreement as a preference signal rather than noise.

Cross-Lingual Transfer under Language Imbalance. Multilingual models such as BLOOMZ (Muennighoff et al., 2022) and Aya (Üstün et al., 2024) demonstrate the benefits of transferring information from high-resource to lower-resource languages, but primarily target aggregate multilingual capability rather than conflicting language-specific preferences. Recent work extends cross-lingual transfer to preference optimization, showing that RLHF and DPO can improve multilingual instruction following but remain sensitive to heterogeneous and imbalanced preference data (Dang et al., 2024). Other approaches transfer alignment signals through cross-lingual consistency, English-derived rewards, or language-imbalance-aware objectives (She et al., 2024; Zhao et al., 2025; Yang et al., 2025b;c). These methods are most natural when desired behavior should remain consistent across languages. In contrast, when preferences conflict, cross-lingual transfer can compete with preserving language-specific supervision.

Language Collapse and Output-Language Control. Off-target generation, where models respond in an unintended language, is a longstanding problem in multilingual generation (Zhang et al., 2020), and English fine-tuning can degrade generation in other languages (Vu et al., 2022). Mechanistic work suggests that multilingual generation depends on both shared and language-specific representations, including language-specific neurons and cross-lingual internal representations that influence output behavior (Wang et al., 2025; Schut et al., 2025; Tang et al., 2024; Ravisankar et al., 2026; Liu & Niehues, 2025; Lee et al., 2025; Lim et al., 2025). Recent studies further show that preference optimization can transfer behavior across languages and reshape language-related representations (Li et al., 2024; Zhang et al., 2026; Zhao et al., 2024), while language-consistency objectives can reduce high-resource-language fallback (Tran et al., 2026). We instead study how language imbalance and preference conflict jointly affect output-language behavior, and why conflicting supervision can reduce collapse into the dominant language. More broadly, prior work typically treats cross-lingual agreement as desirable, whereas pluralistic alignment may require preserving disagreement when preferences legitimately diverge.

3 Multilingual Alignment Under Preference Conflicts

In this section, we formalize multilingual preference learning under language imbalance and cross-lingual preference conflicts. We then describe our experimental setup and present the multilingual alignment paradox across models and training data distributions.

3.1 Problem Definition

Multilingual data imbalance. Let $\mathcal{D} = \{(x_j, y_j^+, y_j^-, \ell_j)\}_{j=1}^{N_{\text{pref}}}$ be a multilingual preference dataset, where each example contains a prompt x_j, a preferred response y_j^+, a dispreferred response y_j^-, and a language label $\ell_j \in \mathcal{L}$. Here, $\mathcal{L}$ is the set of m languages in the dataset. For each language ℓ, let $\mathcal{D}_\ell$ denote its subset of examples and $p_\ell = |\mathcal{D}_\ell|/N_{\text{pref}}$ denotes its proportion of the dataset. We define the *dominant language* as $\ell_d \in \arg\max_{\ell \in \mathcal{L}} p_\ell$ and refer to the remaining languages, $\mathcal{L}_{\text{min}} = \mathcal{L} \setminus \{\ell_d\}$, as *minority languages*. For each minority language ℓ, the *imbalance ratio* $r_\ell = |\mathcal{D}_{\ell_d}|/|\mathcal{D}_\ell|$ compares the number of examples in the dominant language with that in language ℓ.

Cross-lingual preference conflicts. Preferences over semantically equivalent responses may differ across languages due to cultural norms or communication styles. Let $\mathcal{Z}_{ij}$ denote the set of matched prompt–response comparisons $z = (x, y_a, y_b)$ in languages ℓ_i and ℓ_j, and let $\text{pref}_\ell(z) \in \{a, b\}$ identify the preferred candidate. We divide $\mathcal{Z}_{ij}$ into two disjoint partitions:

$$\begin{aligned} \mathcal{Z}_{ij}^{\text{shared}} &= \{z \in \mathcal{Z}_{ij} : \text{pref}_{\ell_i}(z) = \text{pref}_{\ell_j}(z)\}, \\ \mathcal{Z}_{ij}^{\text{conflict}} &= \{z \in \mathcal{Z}_{ij} : \text{pref}_{\ell_i}(z) \neq \text{pref}_{\ell_j}(z)\}. \end{aligned} \tag{1}$$

The shared-preference partition captures agreement across languages, while the conflicting-preference partition captures opposing preferences which are language-specific.

Training objective. We consider supervised fine-tuning (SFT) and direct preference optimization (DPO), with their respective loss functions provided in Appendix B. For either method, the pooled multilingual objective can be expressed as

$$\mathcal{L}_{\text{multi}}(\theta) = \sum_{\ell \in \mathcal{L}} q_\ell \, \mathcal{L}_\ell(\theta, x_l), \tag{2}$$

where $\mathcal{L}_\ell(\theta)$ is the corresponding average training loss for language ℓ, and q_ℓ is its effective sampling weight, satisfying $q_\ell \geq 0$ and $\sum_{\ell \in \mathcal{L}} q_\ell = 1$. For DPO with uniform sampling over the pooled preference dataset, $q_\ell = p_\ell$. Language imbalance therefore gives the dominant language greater weight in the aggregate objective. Under cross-lingual conflicts, this weighting can create tension between improving aggregate performance and preserving minority-language preferences.

3.2 Experimental Setup

Dataset and controlled conditions. We construct a multilingual preference dataset from GlobalOpinionQA (Durmus et al., 2023), covering English, Chinese, French, Arabic, and Hindi. Starting from the same underlying comparisons, we create two preference structures. Under *shared preferences*, all five languages use the same preference ordering. Under *conflicting preferences*, the four non-English languages reverse the English ordering while agreeing with one another. We separately vary the English-to-each-minority training ratio over 1:1, 10:1, and 100:1 by holding the English data fixed and subsampling the minority-language data. Evaluation and test sets are fixed across conditions. Full dataset construction details are in Appendix C.

Models and training. We evaluate BLOOMZ-1.1B (Muennighoff et al., 2023), Llama-3.2-1B-Instruct (Touvron et al., 2023), Qwen3-1.7B (Yang et al., 2025a), and Tiny-Aya-Global-3.35B (Salamanca et al., 2026). For each model, preference structure, and language ratio, we perform SFT and DPO using the corresponding SFT checkpoint as the reference policy. (Detailed in Appendix D.)

Evaluation. For each minority-language input, we jointly evaluate preference correctness and output language. Responses are classified as Target-aligned (TA), Target-misaligned (TM), English-aligned (EA), or English-misaligned (EM), according to whether the response is in the target language or English and whether it matches the corresponding preference. For English inputs, target and English coincide, so TA equals EA.

3.3 The Multilingual Alignment Paradox

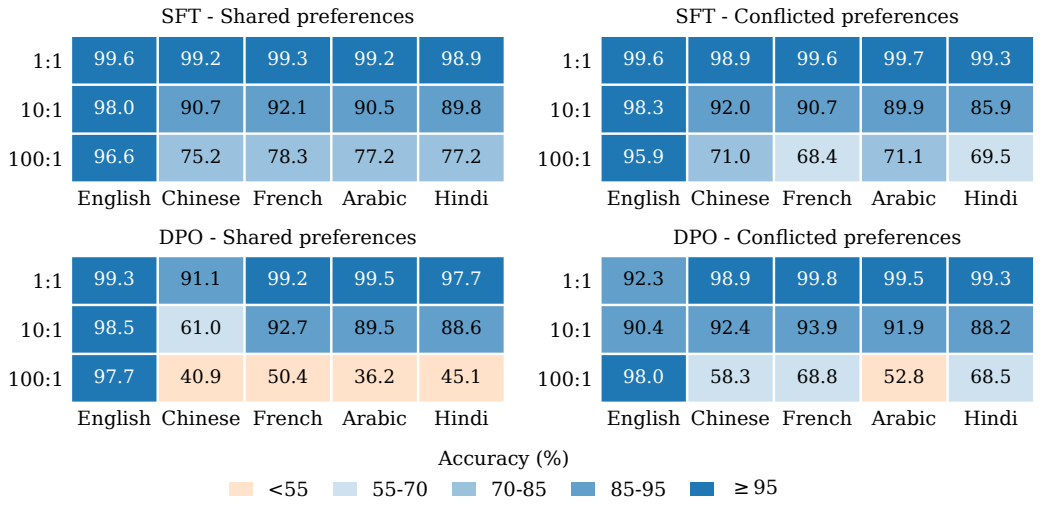


Figure 2: **Model-averaged per-language alignment accuracy under language imbalance and preference conflict.** Accuracy equals TA for non-English inputs and EA for English inputs. Each cell reports the mean percentage averaged equally across 4 models. Within each panel, rows indicate the English-to-minority training-data ratio, and columns indicate the evaluation language. It's shown that alignment with minority preferences worsens as training data imbalance increases.

We compare SFT and DPO across four multilingual models, two preference structures, and three imbalance levels. Under both preference structures, target-language alignment is near perfect at 1:1 but declines with increasing imbalance, most sharply after DPO at 100:1 (Figure 2). SFT outputs remain almost entirely in the target language, whereas DPO increasingly produces English responses. At 100:1, conflict reduces this English-response fraction across all four models.

Alignment with minority preferences worsens as training data imbalance increases. As shown in Figure 2, increasing the English-to-minority training ratio reduces alignment accuracy for target

languages (TA) after both training stages. Averaged across models and minority languages, SFT TA falls from approximately 99% under balanced training to 77.0% with shared preferences and 70.0% with conflicting preferences at 100:1. DPO exhibits larger declines, from 96.9% to 43.2% with shared preferences and from 99.4% to 62.1% with conflicting preferences. At 100:1, English alignment accuracy (EA) remains above 95% in all four model-averaged conditions. These results show that minority-language inputs bear most of the performance loss under imbalance. However, lower TA can reflect either preference errors or language switching.

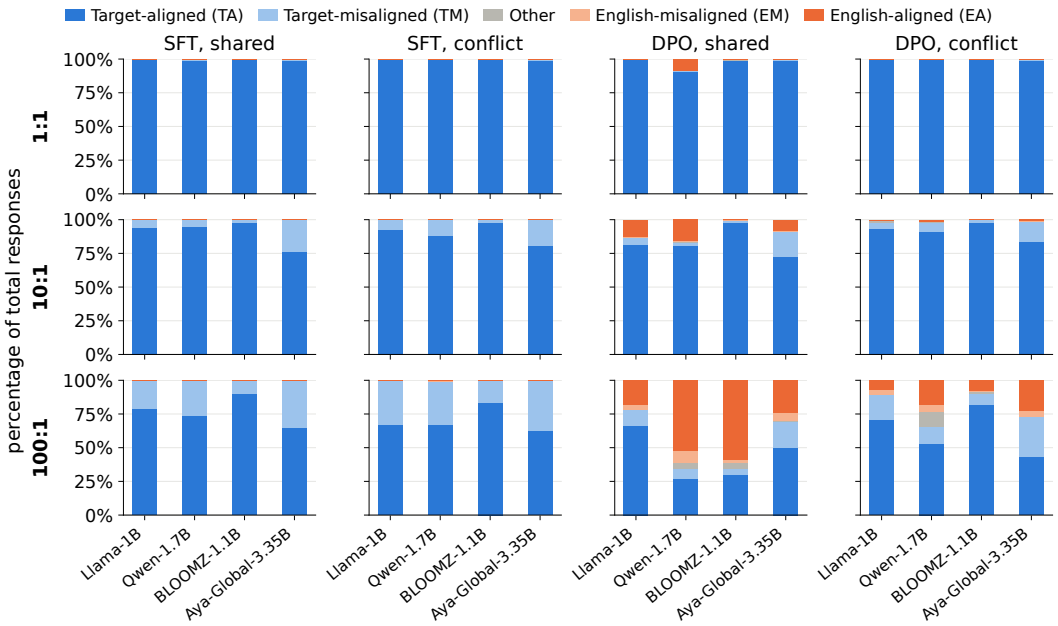


Figure 3: Failure case analysis across models, training stages, preference structures, and language imbalance levels. Rows show different English-to-minority training ratios. The rates are computed by averaging equally over 4 minority languages: Chinese, French, Arabic, and Hindi. It's shown that SFT suffers preference selection, while DPO suffers language collapse, and preference conflicts hurt alignment but prevent language collapse.

SFT suffers preference selection, while DPO suffers language collapse. The error type decomposition of responses in Figure 3 distinguishes these failure modes. After SFT, nearly all responses remain in the target language, with TA + TM close to 100%. As imbalance increases, TA decreases while TM grows, indicating that the principal failure is preference selection within the input language. After DPO, English outcomes become a substantial additional source of error, particularly under shared preferences. At 100:1, English-aligned responses dominate the Qwen3-1.7B and BLOOMZ-1.1B distributions in the shared-preference setting. Because English and target-language preferences are shared in this setting, these EA outcomes express the desired preference but fail to preserve the input language. DPO's lower TA therefore contains both preference errors and language switching.

Preference conflict hurts alignment but prevents language collapse. Comparing preference structures reveals the paradox. After SFT at 100:1, conflicting preferences reduce mean TA from 77.0% to 70.0%, while target-language retention remains nearly complete. Conflict therefore makes preference selection harder in this setting. After DPO, however, conflicting preferences reduce English-response frequency relative to shared preferences across all four models at 100:1 (Figure 3). This improvement in language retention accompanies an increase in mean TA from 43.2% to 62.1%, although substantial TM remains and English responses are not eliminated. Thus, conflicts can hinder preference selection while mitigating language collapse, allowing joint language and preference success to improve after DPO. This benefit is not universal across models, as Aya-Global-3.35B retains lower TA under conflict at 100:1. The central finding is therefore a tension between preference alignment and language preservation, rather than a uniform benefit or cost of conflict.

4 WHY DOES THE MULTILINGUAL ALIGNMENT PARADOX ARISE?

In this section, we explain the paradox through DPO objective analysis and training dynamics (§4.1), followed by representation analysis and causal interventions (§4.2).

4.1 WHAT THE DPO OBJECTIVE CONSTRAINS

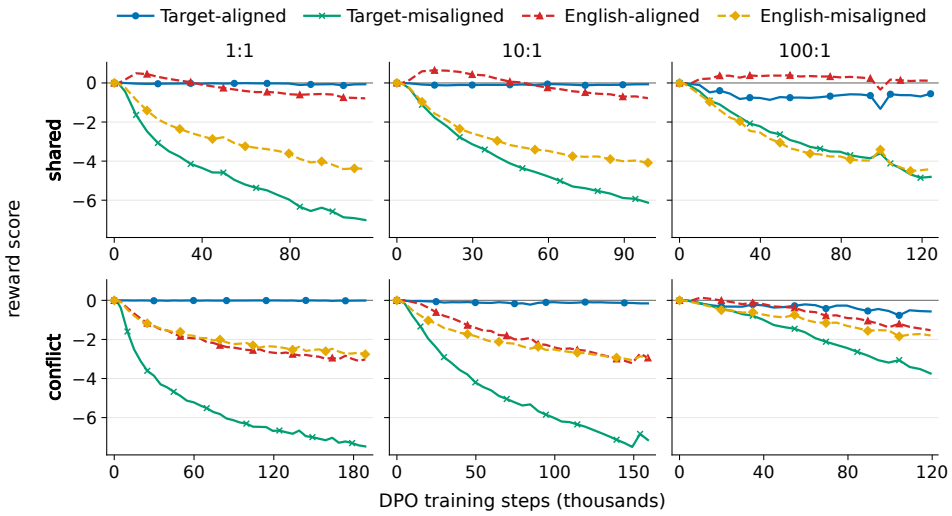


Figure 4: **Reward scores for English and minority-language responses during DPO training of Bloomz-1.1B**. On minority-language prompts with shared preferences at 100:1, English-aligned responses receive higher mean implicit rewards than their target-language counterparts.

DPO under language imbalance. Using the language sampling weights q_ℓ defined in Section §3.1, we specialize the pooled multilingual objective to DPO. For a fixed SFT reference policy π_{ref} and $\beta > 0$, define the reference-adjusted preference margin

$$z_\theta(x, y^+, y^-) = \beta \left[\log \frac{\pi_\theta(y^+ \mid x)}{\pi_{\text{ref}}(y^+ \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_{\text{ref}}(y^- \mid x)} \right]. \tag{3}$$

The multilingual DPO objective is

$$\mathcal{L}_{\text{DPO}}(\theta) = \sum_{\ell \in \mathcal{L}} q_\ell \mathbb{E}_{(x,y^+,y^-)\sim \mathcal{D}_\ell} \left[-\log \sigma\big(z_\theta(x, y^+, y^-)\big) \right]. \tag{4}$$

Within-language comparisons leave language mass unidentified. Partition complete responses into language classes $\{\mathcal{Y}_j : j \in \mathcal{J}\}$. Define the *language-selection probability* $\rho_\theta(j \mid x)$ and the *conditional response-generation probability* $p_\theta(y \mid x, j)$ as

$$\rho_\theta(j \mid x) = \sum_{y \in \mathcal{Y}_j} \pi_\theta(y \mid x), \qquad p_\theta(y \mid x, j) = \frac{\pi_\theta(y \mid x)}{\rho_\theta(j \mid x)} \quad (y \in \mathcal{Y}_j). \tag{5}$$

The former measures total probability assigned to language j, and the latter measures the probability of a particular response conditional on that language. Thus, $\pi_\theta(y \mid x) = \rho_\theta(j \mid x) p_\theta(y \mid x, j)$. This is a probabilistic decomposition, not an architectural assumption.

Both responses in each training pair use the input language ℓ, so their language-selection factors cancel under both the policy and reference in DPO loss:

$$z_\theta(x, y^+, y^-) = \beta \left[\log \frac{p_\theta(y^+ \mid x, \ell)}{p_\theta(y^- \mid x, \ell)} - \log \frac{p_{\text{ref}}(y^+ \mid x, \ell)}{p_{\text{ref}}(y^- \mid x, \ell)} \right]. \tag{6}$$

Holding the within-language conditional distributions fixed, valid changes to ρ_θ therefore leave the DPO loss unchanged. In contrast, SFT explicitly penalizes low target-language probability:

$$-\log \pi_\theta(y^+ \mid x) = -\log \rho_\theta(\ell \mid x) - \log p_\theta(y^+ \mid x, \ell). \tag{7}$$

This distinction explains the *stronger language preservation* after SFT than DPO.

Implicit reward gaps reveal shifts toward English responses. For a minority-language input x_ℓ, let y_{EN} and y_ℓ express the same answer in English and the input language. Define the DPO implicit reward and its cross-language gap:

$$\hat{r}_\theta(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\mathrm{ref}}(y \mid x)}, \qquad G_\theta(x_\ell) = \hat{r}_\theta(x_\ell, y_{\mathrm{EN}}) - \hat{r}_\theta(x_\ell, y_\ell). \tag{8}$$

Writing $\Delta_{\mathrm{ref}} F = F_\theta - F_{\mathrm{ref}}$, the language–response decomposition gives

$$\begin{aligned} \frac{G_\theta(x_\ell)}{\beta} = \Delta_{\mathrm{ref}} \log \frac{\rho(\mathrm{EN} \mid x_\ell)}{\rho(\ell \mid x_\ell)} \\ + \Delta_{\mathrm{ref}} \log \frac{p(y_{\mathrm{EN}} \mid x_\ell, \mathrm{EN})}{p(y_\ell \mid x_\ell, \ell)}. \end{aligned} \tag{9}$$

Thus, even for semantically matched responses, the gap reflects both language probabilities and conditional response probabilities. Equivalently,

$$\log \frac{\pi_\theta(y_{\mathrm{EN}} \mid x_\ell)}{\pi_\theta(y_\ell \mid x_\ell)} = \log \frac{\pi_{\mathrm{ref}}(y_{\mathrm{EN}} \mid x_\ell)}{\pi_{\mathrm{ref}}(y_\ell \mid x_\ell)} + \frac{G_\theta(x_\ell)}{\beta}. \tag{10}$$

Figure 4 plots the mean implicit rewards for target-language and English responses to target-language inputs, averaged across the four minority languages. Under shared preferences at 100:1 imbalance, BLOOMZ-1.1B assigns higher mean implicit rewards to preferred English responses than to their target-language counterparts. This positive mean gap ($\mathbb{E}x\ell[G_\theta(x_\ell)]/\beta > 0$) indicates a relative shift toward English for these matched responses, consistent with increased English fallback. The other three models also exhibit positive mean gaps under the same setting (shown in Appendix F).

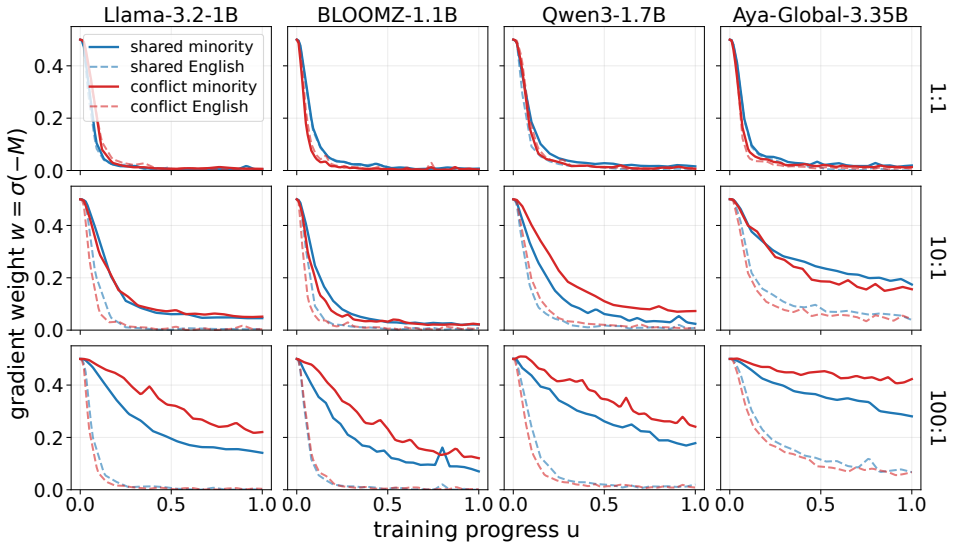


Figure 5: **DPO preference weights during training.** Mean per-example DPO weight $w = \sigma(-z)$ for English and minority-language examples under shared and conflicting preferences. Columns show models and rows show English-to-each-minority training ratios; training progress is normalized to $[0, 1]$. Under severe imbalance, minority-language weights remain higher under conflict, while English weights rapidly approach zero, indicating stronger residual minority preference signals.

Preference conflict strengthens local minority corrections. Consider a two-coordinate surrogate with an English-to-minority sampling ratio R:1. For $j \in \{\mathrm{EN}, \ell\}$, define the reference-adjusted log-probability ratio

$$d_j(y) = \log \frac{\pi_\theta(y \mid x_j)}{\pi_{\mathrm{ref}}(y \mid x_j)}, \qquad a = \beta[d_{\mathrm{EN}}(y_{\mathrm{EN}}^+) - d_{\mathrm{EN}}(y_{\mathrm{EN}}^-)], \qquad b = \beta[d_\ell(y_\ell^+) - d_\ell(y_\ell^-)] - sa, \tag{11}$$

where y_j^+ and y_j^- follow language j's preferences, and $s = +1$ for shared preferences and $s = -1$ for conflict. Thus, a is the English DPO margin, while b is the residual minority margin after subtracting the assumed shared contribution sa. The surrogate treats a and b as independently adjustable coordinates:

$$z_{\mathrm{EN}} = a, \qquad z_\ell = sa + b. \tag{12}$$

Increasing a helps the minority margin under agreement but hurts it under conflict; increasing b improves the minority margin without changing the English margin. These coordinates describe preference margins, not output-language probabilities.

With $\phi(z) = -\log \sigma(z)$, the sampling-weighted loss is

$$\mathcal{L}_s = \frac{R\phi(a) + \phi(sa + b)}{R + 1}. \tag{13}$$

Since $\phi'(z) = -\sigma(-z)$, smaller margins yield stronger corrective signals. Euclidean gradient descent in these coordinates, with step size η and $c = \eta/(R + 1)$, gives

$$\Delta a_s = c[R\sigma(-a) + s\sigma(-sa - b)], \qquad \Delta b_s = c\sigma(-sa - b). \tag{14}$$

At a common state with $a > 0$ in imbalanced settings (shown in Figure 4),

$$\Delta b_{\mathrm{conflict}} = c\sigma(a - b) > c\sigma(-a - b) = \Delta b_{\mathrm{shared}}. \tag{15}$$

Thus, English-favored progress reduces the residual minority signal under shared preferences, whereas conflict sustains a stronger local minority correction. To examine this mechanism empirically, we track the mean per-example DPO weight $w_\ell = \mathbb{E}_{i\sim\mathcal{D}_\ell}[\sigma(-z_i)]$, which scales preference updates before accounting for gradient direction and magnitude. In the surrogate, this weight directly determines the minority-specific update, $\Delta b = cw_\ell$. Under 100:1 imbalance, minority-language weights remain higher under conflicting than shared preferences through most of training across all four models, while English weights rapidly approach zero (Figure 5). These observations support sustained minority preference signals under conflict, which can preserve language-specific representations supporting target-language generation (§4.2).

4.2 Representational Evidence and Causal Interventions

The analysis in §4.1 suggests that conflicting preferences can sustain language-conditioned corrections that are not satisfied by progress along a shared dominant-language direction. We next ask whether this difference is reflected in the representations learned under shared and conflicting preferences, and whether language-associated representations influence English fallback.

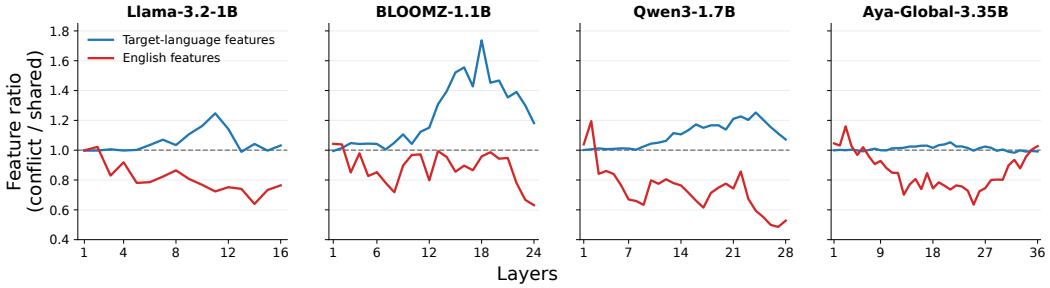


Figure 6: **Language-associated SAE activity under shared and conflicting preferences at 100:1 imbalance.** For each decoder layer, curves show the conflict-to-shared ratio of active target-language-associated and English-associated features, averaged equally across the four minority languages. Values above one indicate greater activity under conflict. Conflict generally reduces English-associated activity across all four models, while higher target-language-associated activity is pronounced for most models. Feature counts are computed per prompt from a TopK SAE.

Conflict shifts representations away from English and toward the target language. We analyze residual-stream activations using TopK sparse autoencoders (SAEs) (Makhzani & Frey, 2013; Gao et al., 2025). For each model and layer, we train a common SAE on activations from both the shared and conflicting preference checkpoints, providing a shared feature basis for comparing the

Model	Layer (Total)	d_ℓ	Fallback (%)			
			Baseline	Amplified	Δ (pp)	Δ PC (pp)
Shared preferences						
Llama-3.2-1B	13 (16)	0.40	21.7	1.9	−19.8	−1.7
BLOOMZ-1.1B	20 (24)	0.60	60.7	0.0	−60.7	−0.9
Qwen3-1.7B	24 (28)	0.60	62.4	20.3	−42.0	−0.8
Tiny-Aya-Global-3.35B	31 (36)	1.60	33.1	1.2	−31.9	−1.8
Conflicting preferences						
Llama-3.2-1B	13 (16)	1.25	9.7	0.4	−9.3	−1.4
BLOOMZ-1.1B	20 (24)	0.90	9.4	0.0	−9.3	+2.6
Qwen3-1.7B	24 (28)	1.25	22.6	1.9	−20.7	+3.9
Tiny-Aya-Global-3.35B	31 (36)	1.25	27.8	3.6	−24.2	+14.2

Table 1: **Effect of amplifying target-language-associated SAE features on English fallback.** For each model and preference condition, we intervene at the indicated layer of the 100:1 DPO checkpoint. d_ℓ is the normalized intervention dose. Δ denotes intervention minus baseline in percentage points; PC denotes preference correctness.

two conditions. We identify features associated with English and with each target language, and measure their activity on minority-language prompts. Full SAE training and feature-labeling details are provided in Appendix G. Figure 6 shows a consistent shift in language-associated activity under conflicting preferences. English-associated activity is lower in all models, while target-language activity is generally higher under conflict. Thus, the reduction in English fallback under conflicting preferences is accompanied by an internal representational shift away from the dominant language and toward the target language. These comparisons are descriptive, however, and do not establish that the identified features causally control output language.

Amplifying target-language-associated features reduces English fallback. We thus intervene directly on target-language-associated representations. For each minority language and layer, we identify a set of SAE features whose activations predict the prompt language using probes fitted on our training data. At inference time, we amplify the corresponding feature directions at the final prompt token and evaluate the resulting model behavior on minority-language prompts from the 100:1 checkpoints. We test multiple layers and intervention scales (Appendix G). Table 1 reports one setting per model and preference condition. Amplifying target-language-associated features reduces English fallback across all four models and both preference conditions. Preference correctness changes little in most settings and improves in several conflicting-preference checkpoints. This provides causal evidence that the target-language-associated representations that differ between the two training conditions are behaviorally relevant to output-language selection. A complementary intervention shows the same trend from the opposite direction: removing English-associated SAE features also reduces English fallback while largely preserving preference correctness (Table 5, Appendix G).

Together, these results show that conflicting supervision is accompanied by weaker English-associated and stronger target-language-associated representations, and that manipulating these language-associated representations changes the model's tendency to fall back to English. Combined with the optimization analysis in §4.1, this supports the interpretation that preference conflict sustains language-conditioned representations that help preserve the target language under severe imbalance.

5 Conclusions

We identify a multilingual alignment paradox: under language imbalance, cross-lingual preference conflict can hinder preference selection yet prevent language collapse. Across four models and five languages, SFT largely preserves the input language despite preference errors, whereas DPO increasingly switches to the dominant language, English, especially under shared preferences. Objective analysis and training dynamics suggest that conflict sustains minority preference signals, while SAE interventions show that language-associated representations influence English fallback. Together, these findings highlight the need to preserve both language-specific preferences and output language in the multilingual alignment objective design.

AI USAGE STATEMENT

We disclose the following. All core contributions of this work, including the problem formulation, experimental design, and the analysis were conceived and developed entirely by the authors. Large language models (e.g., ChatGPT, Claude) were used only as assistive tools in this work. First, as described in Appendix C, GPT-4o-mini was used to translate dataset fields from English into target languages during dataset construction. Second, LLM tools were used to help draft portions of experiment code, plotting scripts, and manuscript text, as well as for language polishing and editing. In all cases, the LLM-generated outputs were reviewed, verified, and revised by the authors to ensure correctness and faithfulness to the intended content. No LLM was used to produce novel scientific ideas and design experiments.

REFERENCES

Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. Rlhf can speak many languages: Unlocking multilingual preference optimization for llms. *arXiv preprint arXiv:2407.02552*, 2024.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. Towards measuring the representation of subjective global opinions in language models, 2023.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pp. 26721–26754, 2025.

HyoJung Han, Sweta Agrawal, and Eleftheria Briakou. Rethinking cross-lingual alignment: Balancing transfer and cultural erasure in multilingual llms. *arXiv preprint arXiv:2510.26024*, 2025.

Jungseob Lee, Seongtae Hong, Hyeonseok Moon, and Heui-Seok Lim. Cross-lingual optimization for language transfer in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15100–15119, 2025.

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 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.784. URL https://aclanthology.org/2024.findings-emnlp.784/.

Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. Language-specific latent process hinders cross-lingual performance. *arXiv preprint arXiv:2505.13141*, 2025.

Danni Liu and Jan Niehues. Middle-layer representation alignment for cross-lingual transfer in fine-tuned llms. *arXiv preprint arXiv:2502.14830*, 2025.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*, 2022.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st annual meeting of the Association for Computational Linguistics (volume 1: Long papers)*, pp. 15991–16111, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. A survey of multilingual large language models. *Patterns*, 6(1), 2025.

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

Kartik Ravisankar, Hyojung Han, Sarah Wiegreffe, and Marine Carpuat. Can you map it to english? the role of cross-lingual alignment in the multilingual performance of llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4854–4872, 2026.

Michael J Ryan, William Held, and Diyi Yang. Unintended impacts of llm alignment on global representation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16121–16140, 2024.

Alejandro R Salamanca, Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaee, Mehrnaz Mofakhami, Ananya Sahu, et al. Tiny aya: Bridging scale and multilingual depth. *arXiv preprint arXiv:2603.11510*, 2026.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*, 2025.

Shuaijie She, Shujian Huang, Wei Zou, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *arXiv preprint arXiv:2401.06838*, 2024.

Kazuhiro Takemoto. The moral machine experiment on large language models. *Royal Society open science*, 11(2), 2024.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5701–5715, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. LaCoMSA: Language-consistency multilingual self-alignment with latent representation rewarding. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4839–4853, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.224. URL https://aclanthology.org/2026.eacl-long.224/.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*, 2024.

Tu Vu, Aditya Barua, Brian Lester, Daniel Cer, Mohit Iyyer, and Noah Constant. Overcoming catastrophic forgetting in zero-shot cross-lingual generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9279–9300, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.630. URL https://aclanthology.org/2022.emnlp-main.630/.

W3Techs. Usage statistics of content languages for websites. https://w3techs.com/technologies/overview/content_language, 2026. Accessed: 2026-04-01.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5075–5094, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.253. URL https://aclanthology.org/2025.acl-long.253/.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. *arXiv preprint arXiv:2411.04986*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. Implicit cross-lingual rewarding for efficient multilingual preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21125–21147, 2025b.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. Language imbalance driven rewarding for multilingual self-improving. In *International Conference on Learning Representations*, volume 2025, pp. 43911–43936, 2025c.

Binwei Yao, Zefan Cai, Yun-Shiuan Chuang, Shanglin Yang, Ming Jiang, Diyi Yang, and Junjie Hu. No preference left behind: Group distributional preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.148. URL https://aclanthology.org/2020.acl-main.148/.

Shimao Zhang, Zhejian Lai, Xiang Liu, Shuaijie She, Xiao Liu, Yeyun Gong, Shujian Huang, and Jiajun Chen. How does alignment enhance llms' multilingual capabilities? a language neurons perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34799–34807, 2026.

Weixiang Zhao, Yulin Hu, Jiahe Guo, Xingyu Sui, Tongtong Wu, Yang Deng, Yanyan Zhao, Bing Qin, Wanxiang Che, and Ting Liu. Lens: Rethinking multilingual enhancement for large language models. *arXiv preprint arXiv:2410.04407*, 2024.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, et al. Mpo: Multilingual safety alignment via reward gap optimization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23564–23587, 2025.

A Limitations and Implications

Our study uses a controlled multilingual preference setting derived from GlobalOpinionQA, where the underlying comparisons are fixed while preference structure and language imbalance are varied. This design isolates the effects of conflict and imbalance, but naturally occurring multilingual preferences may be noisier, more partial, or more open-ended. We evaluate four multilingual models, five languages, and SFT followed by DPO with English as the dominant language, so future work should test larger models, other dominant languages, and other alignment objectives. Our SAE interventions show that language-associated representations are behaviorally relevant to English fallback, but do not establish that they fully mediate the effect of preference conflict.

Our results also suggest that multilingual alignment should evaluate preference correctness and output language separately. A model can express the preferred answer while switching into the dominant language, so preference accuracy alone can hide language collapse. More broadly, cross-lingual transfer is not always beneficial: when preferences are shared, dominant-language transfer can help propagate desired behavior, but when preferences differ, the same pressure can suppress language-specific signals. Alignment methods should therefore account for both preference pluralism and language preservation rather than treating cross-lingual agreement as a universal goal.

B Loss Functions of Preference Learning

To isolate the effect of preference training from pretraining biases, we fine-tune models on the multilingual conflict preference dataset in two stages: (1) Supervised Fine-Tuning (SFT) and (2) Direct Preference Optimization (DPO).

Stage 1: Supervised Fine-Tuning. Given a pretrained language model with parameters θ_0, we first perform supervised fine-tuning on the instruction-following data to adapt the model to the task distribution:

$$\mathcal{L}_{\text{SFT}}(\theta; x, y) = -\log \pi_\theta(y|x). \tag{16}$$

The SFT objective aggregates across languages: $\mathcal{L}_{\text{SFT}}^{\text{total}} = \mathbb{E}_{\ell\sim p(\ell)}[\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{SFT}}^{(\ell)}}[\mathcal{L}_{\text{SFT}}]]$, producing an SFT model θ_{SFT} that serves as both the initialization for preference learning and the reference policy $\pi_{\text{ref}} = \pi_{\theta_{\text{SFT}}}$.

Stage 2: Direct Preference Optimization. Starting from θ_{SFT}, we optimize the Direct Preference Optimization (DPO) objective (Rafailov et al., 2023):

$$\mathcal{L}_{\text{DPO}}(\theta; x, y^+, y^-) = -\log \sigma\left(\beta \log \frac{\pi_\theta(y^+|x)}{\pi_{\text{ref}}(y^+|x)} - \beta \log \frac{\pi_\theta(y^-|x)}{\pi_{\text{ref}}(y^-|x)}\right), \tag{17}$$

where $\sigma(\cdot)$ is the sigmoid function and β is the temperature parameter. Standard multilingual alignment aggregates losses across languages: $\mathcal{L}_{\text{total}} = \mathbb{E}_{\ell\sim p(\ell)}[\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}_{\text{pref}}^{(\ell)}}[\mathcal{L}_{\text{DPO}}]]$. However, this approach often results in dominant languages overwhelming minority language preferences during optimization, particularly problematic when preference conflicts exist, as the learned preference model implicitly encodes a biased compromise that favors high-resource languages.

C Dataset Construction and Processing Pipeline

Data Statistics Before downsampling, each language contains 14,350 training, 1,794 validation, and 1,794 test prompts, aligned across the five languages. Validation and test prompts remain fixed across settings (8,970 prompts per split). English retains all 14,350 training prompts, while each of Chinese, French, Arabic, and Hindi retains 14,350, 1,435, or 144 at nominal English-to-each-minority ratios of 1:1, 10:1, or 100:1, respectively. The corresponding training totals are 71,750, 20,090, and 14,926. Shared and conflicting settings have identical sample counts.

Our experiments are built on a controlled synthetic multilingual preference dataset derived from the Anthropic/`llm_global_opinions` corpus Durmus et al. (2023), which contains survey-style opinion questions with associated response distributions reflecting public attitudes across diverse

topics. The original release is English-only and is not directly organized for our pairwise preference learning setting. We therefore apply a multi-stage processing pipeline, illustrated in Figure 7, to construct the final multilingual preference datasets used in our experiments.

Raw Data and Question Selection We first identify the subset of questions shared across the English, Chinese, and French settings used in this paper. We then reorganize the data by question, such that each question forms a group containing multiple stylistically varied samples. This grouped structure serves as the basis for all subsequent multilingual controls.

Translation and Alignment Since the source data is exclusively in English, we translate the relevant fields, including prompts, accepted and rejected responses, answer options, and choice labels into Chinese, French, Hindi and Arabic using GPT-4o-mini. To ensure translation quality, we employ a two-pass strategy: an initial parallel translation pass followed by a targeted re-translation pass that identifies and retranslates any entries that failed in the first round. This produces five parallel datasets (English, Chinese, French, Hindi and Arabic) that share the same underlying question structure. During this step, we preserve question identity and within-question sample alignment across languages. This design ensures that downstream comparisons across languages are made on semantically matched question sets, rather than being confounded by differences in topic coverage or prompt format.

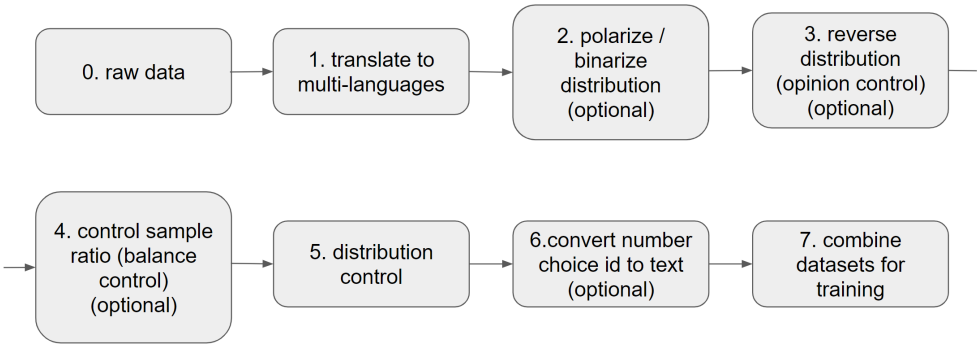


Figure 7: Overview of the dataset construction and processing pipeline. Starting from the raw English source data, we select the shared multilingual question set, translate aligned fields into other languages, optionally binarize and polarize the preference distribution, control cross-lingual opinion conflict and language balance, and finally convert grouped samples into training-ready preference pairs.

Binary Reduction and Polarization To facilitate controlled comparison of multilingual preference conflicts, we optionally reduce the original multi-way answer space to a binary decision space. In this step, semantically related options are collapsed into two poles, yielding a cleaner positive/negative-style representation that removes variation caused by heterogeneous answer inventories across questions. To construct a clearer supervision signal for each language, we further build polarized versions of the dataset on top of this binary form by concentrating the answer distribution of each question onto a single dominant pole. This makes preference conflicts across languages more directly comparable and provides a controlled setting for testing whether multilingual models preserve language-specific preferences or collapse toward the dominant-language preference.

Opinion Control and Language Balance We further introduce two orthogonal controls over the resulting dataset. The first is *opinion control*, which uses a *reverse rate* to adjust the proportion of questions on which a target language agrees with the English preference distribution. At one extreme, the target language matches English on all shared questions; at the other extreme, it adopts the fully reversed preference distribution; intermediate settings mix matched and reversed questions. This provides a direct handle on the degree of cross-lingual preference conflict while keeping the semantic content fixed.

The second is *balance control*, which adjusts the amount of non-English training data relative to English via a *sample ratio* parameter. We keep the English training set fixed and vary the number of training samples in each of Chinese, French, Arabic, and Hindi to simulate settings ranging from balanced multilingual supervision to highly skewed resource imbalance. This allows us to simulate realistic data imbalance. Evaluation and test sets are held fixed across all ratio settings to ensure fair comparison.

Conversion to Training Pairs After applying the above transformations, we construct the final preference pairs for training. For each question group, samples are assigned accepted and rejected responses according to the controlled opinion distribution, forming the preference pairs which are then split into train, evaluation, and test sets for supervised fine-tuning and preference optimization. In addition to the main construction above, we also convert the numeric choice identifiers into language-specific textual labels to examine whether surface-level label representation affects model behavior. In some experimental configurations, we also consider an alternative prompt format that presents the options as explicit A/B choices within the prompt text itself to test whether prompt surface form affects model behavior.

D Experimental Settings

Models and training. We evaluate four multilingual models: BLOOMZ-1.1B, Llama-3.2-1B-Instruct, Qwen3-1.7B, and Tiny-Aya-Global-3.35B. For each model, we train under both shared and conflicting preferences and English-to-each-minority ratios of 1:1, 10:1, and 100:1. We first perform supervised fine-tuning (SFT), followed by Direct Preference Optimization (DPO) using the corresponding SFT checkpoint as the reference policy. We set $\beta = 0.1$ for DPO and report checkpoints selected by validation performance.

We train each model with SFT and then DPO, using the SFT checkpoint as the reference policy. Both stages use RMSprop with a learning rate of 5×10^{-7}, 150 warmup steps, no gradient accumulation, and a maximum length of 512 tokens (256 for the prompt). The global batch size is 144 for SFT and 72 preference pairs for DPO for the 1B-scale models, and 192 and 96 for Aya-Global-3.35B. Training runs for at most 15 epochs with early stopping: we evaluate every $\sim$10,000 training examples in SFT and every $\sim$5,000 preference pairs in DPO, stop after 5 (SFT) or 8 (DPO) evaluations without improvement, and report the checkpoint with the lowest validation loss. All runs use FSDP with bfloat16 mixed precision.

Across all experiments, we use FSDP and set the computation data type to `bfloat16`. We disable sampling during evaluation in both SFT and DPO. We report the performance of the checkpoint with the best validation performance.

Evaluation metrics. We jointly evaluate response language and preference alignment. For each non-English input language ℓ, we classify responses into four categories: 1) *Target-aligned (TA)*, responses in ℓ that agree with ℓ's preferences; 2) *Target-misaligned (TM)*, responses in ℓ that disagree with ℓ's preferences; 3) *English-aligned (EA)*, responses in English that agree with English preferences; and 4) *English-misaligned (EM)*, responses in English that disagree with English preferences. We report each category as a percentage of all evaluation examples in input language ℓ. For English inputs, the target and reference languages coincide, so we report only two English rates.

E Detailed Accuracy Across Models

Model	Stage	Setting	Ratio	EN	ZH	FR	AR	HI	Mean
BLOOMZ-1.1B	SFT	Shared	1:1	99.22	99.61	100.00	98.83	100.00	99.53
		Shared	10:1	100.00	97.27	98.44	98.05	98.05	98.36
		Shared	100:1	99.61	88.28	93.75	91.02	88.28	92.19
		Conflicted	1:1	100.00	99.22	100.00	100.00	100.00	99.84
		Conflicted	10:1	99.22	98.05	96.88	97.66	97.27	97.81
		Conflicted	100:1	100.00	84.77	83.98	81.25	82.42	86.48
	DPO	Shared	1:1	100.00	99.61	99.61	99.61	96.48	99.06
		Shared	10:1	99.61	97.66	97.66	97.66	97.66	98.05
		Shared	100:1	100.00	78.13	10.16	20.70	12.11	44.22
		Conflicted	1:1	99.61	99.61	100.00	100.00	100.00	99.84
		Conflicted	10:1	99.61	97.27	98.44	98.44	97.66	98.28
		Conflicted	100:1	99.22	74.61	83.59	85.55	82.42	85.08
Llama-1B	SFT	Shared	1:1	100.00	99.22	100.00	99.61	100.00	99.77
		Shared	10:1	99.61	93.36	95.70	92.97	94.92	95.31
		Shared	100:1	99.61	77.73	83.59	75.78	79.30	83.20
		Conflicted	1:1	100.00	98.83	99.61	99.22	100.00	99.53
		Conflicted	10:1	99.61	94.92	91.80	92.19	90.63	93.83
		Conflicted	100:1	97.27	69.14	61.72	69.92	66.80	72.97
	DPO	Shared	1:1	99.61	99.22	100.00	99.61	100.00	99.69
		Shared	10:1	99.61	57.03	92.58	90.23	85.16	84.92
		Shared	100:1	99.22	48.83	83.59	76.56	55.47	72.73
		Conflicted	1:1	100.00	98.44	100.00	99.22	99.61	99.45
		Conflicted	10:1	100.00	92.58	95.70	94.53	91.41	94.84
		Conflicted	100:1	99.22	68.36	75.78	71.48	67.58	76.48
Qwen3-1.7B	SFT	Shared	1:1	99.61	98.83	98.44	99.22	98.44	98.91
		Shared	10:1	98.05	96.88	94.14	94.14	92.58	95.16
		Shared	100:1	98.83	73.83	72.66	74.61	73.44	78.67
		Conflicted	1:1	99.22	98.83	100.00	100.00	99.61	99.53
		Conflicted	10:1	98.05	91.02	89.45	89.06	83.98	90.31
		Conflicted	100:1	97.66	69.14	68.75	68.75	62.11	73.28
	DPO	Shared	1:1	97.66	66.41	98.44	99.22	97.27	91.80
		Shared	10:1	99.61	38.28	97.27	90.63	96.09	84.38
		Shared	100:1	99.61	16.02	37.89	5.47	46.48	41.09
		Conflicted	1:1	100.00	99.22	100.00	99.61	99.22	99.61
		Conflicted	10:1	98.83	92.97	93.75	88.67	89.84	92.81
		Conflicted	100:1	99.61	51.56	69.14	31.25	58.98	62.11
Aya-Global-3.35B	SFT	Shared	1:1	99.61	99.22	98.83	99.22	97.27	98.83
		Shared	10:1	94.14	75.39	80.08	76.95	73.44	80.00
		Shared	100:1	88.28	60.94	63.28	67.19	67.58	69.45
		Conflicted	1:1	99.22	98.83	98.83	99.61	97.66	98.83
		Conflicted	10:1	96.48	83.98	84.77	80.86	71.88	83.59
		Conflicted	100:1	88.67	60.94	58.98	64.45	66.80	67.97
	DPO	Shared	1:1	100.00	99.22	98.83	99.61	96.88	98.91
		Shared	10:1	95.31	51.17	83.20	79.69	75.39	76.95
		Shared	100:1	91.80	20.70	69.92	42.19	66.41	58.20
		Conflicted	1:1	69.53	98.44	99.22	99.22	98.44	92.97
		Conflicted	10:1	63.28	86.72	87.89	85.94	73.83	79.53
		Conflicted	100:1	94.14	38.67	46.88	23.05	65.23	53.59

Table 2: Per-language preference accuracy (%) by model, training stage, preference setting, and English:minority data ratio. Higher is better.

F Reward Scores for English and Minority-language Responses during DPO Training

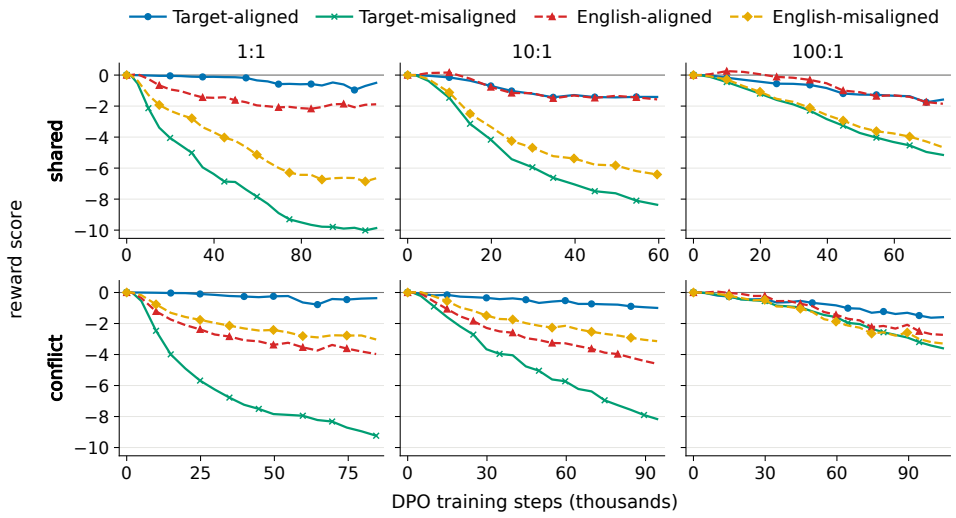


Figure 8: **Reward scores for English and minority-language responses during DPO training of Llama-1B**. On minority-language prompts with shared preferences at 100:1, English-aligned responses receive slightly higher mean implicit rewards than target-aligned responses for most of training, but the gap peaks early and closes by the end; this English advantage is weaker than for BLOOMZ-1.1B and Qwen-1.7B. With conflicting preferences, English-aligned responses end well below target-aligned ones at every ratio.

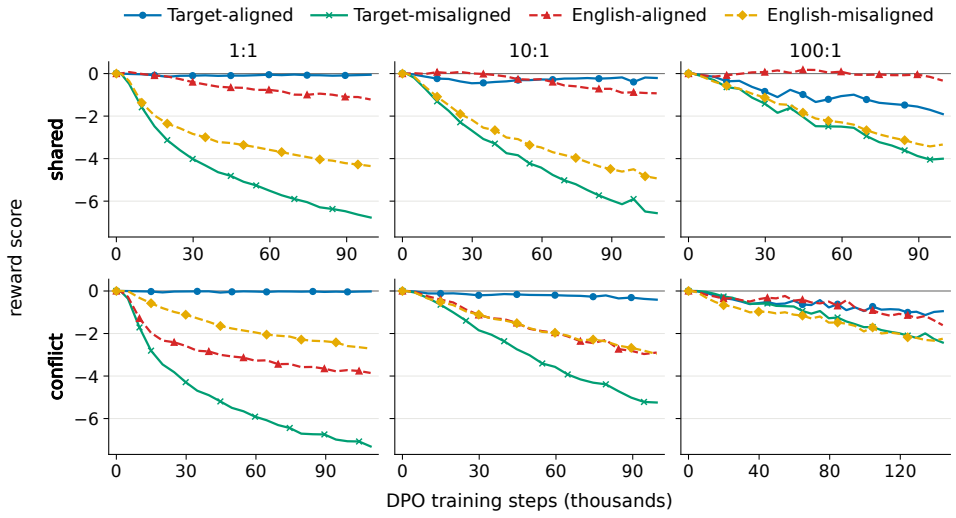


Figure 9: **Reward scores for English and minority-language responses during DPO training of Qwen-1.7B**. On minority-language prompts with shared preferences at 100:1, English-aligned responses stay near zero while target-aligned responses decline, so English-aligned responses receive higher mean implicit rewards from early in training onward; this gap is the largest among the four models. With conflicting preferences, English-aligned responses stay well below target-aligned ones at 1:1 and 10:1 and end slightly below them at 100:1.

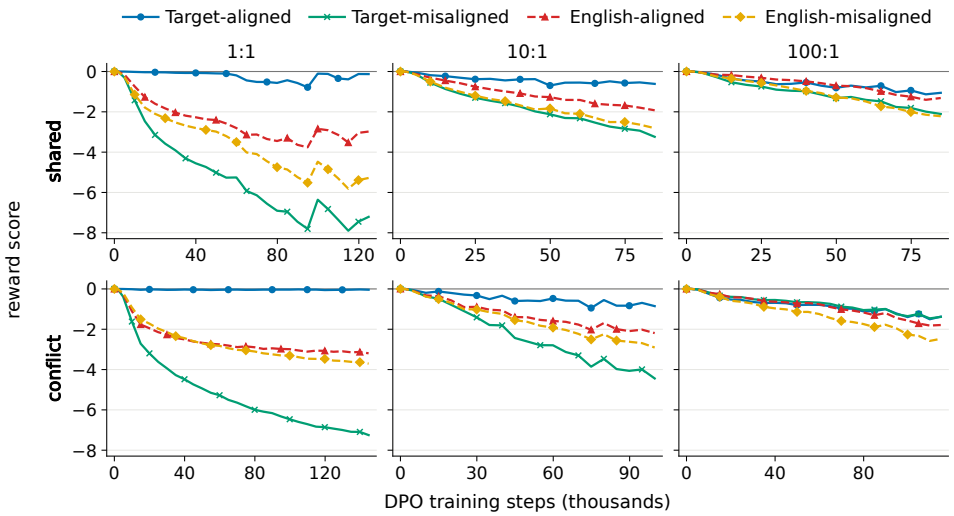


Figure 10: **Reward scores for English and minority-language responses during DPO training of Aya-Global-3.35B**. On minority-language prompts with shared preferences, English-aligned responses end below target-aligned ones at every ratio, but the gap narrows as the English:minority ratio grows and almost closes at 100:1, where English-aligned responses are marginally ahead early in training. With conflicting preferences, English-aligned responses also end below target-aligned ones at every ratio.

G Additional Details on SAE Experiments

This section provides additional details for the SAE analysis in §4.2. We first describe SAE training and the feature-labeling procedure used for Figure 6. We then give the full setup for the target-language amplification and English-feature ablation experiments.

SAE Training. We use TopK sparse autoencoders (SAEs) with 8,192 features and $k = 32$ to analyze residual-stream activations at the final prompt token. We train one SAE for each model and decoder layer. For each model and layer, the SAE is trained jointly on activations from the shared and conflicting preference checkpoints using the same balanced multilingual training prompts. This gives the two conditions the same feature dictionary within each model and layer.

Let h_ℓ denote the residual-stream activation at layer ℓ. The SAE operates on the normalized activation,

$$\tilde{h}_\ell = \frac{h_\ell - \mu_\ell}{\sigma_\ell}$$

where μ_ℓ and σ_ℓ are the layer-specific normalization statistics. All analyses below use activations at the final prompt token unless otherwise stated.

G.1 Identifying Feature Activity in Shared and Conflict Preference Settings

Identifying Language-Associated Features. For Figure 6, we label features using their activation frequencies across languages. Feature labels are computed from evaluation prompts, averaging equally over the shared and conflicting preference checkpoints. For each feature, we sum its positive-activation frequency separately over the five languages. A feature is associated with language L if at least 80% of its total positive-activation frequency comes from prompts in L. Features with zero total frequency or no language above this threshold are left unassigned. For a minority-language prompt in target language X, we count active English-associated features and active features associated with X. We refer to the latter as target-language-associated features. Counts are first computed separately for each minority language and then averaged equally across the four minority languages.

Layerwise Feature Activity. For each layer, we compare language-associated feature activity between the conflicting and shared preference checkpoints. For target-language-associated features, we compute the conflict-to-shared ratio of these language-averaged counts. A ratio above one means that target-language-associated activity is higher under conflict. As shown in Figure 6, target-language-associated activity is generally higher under conflict, with the clearest differences appearing in Llama, BLOOMZ, and Qwen. We use the same comparison for English-associated features. Here, a ratio below one means that English-associated activity is lower under conflict. English-associated activity is lower across most layers and all four models, with larger differences appearing in the middle and later layers.

G.2 Target-Language Feature Amplification

Feature selection. For each layer ℓ and minority language X, we fit an ℓ_1-regularized logistic regression probe with $C = 0.1$ on training prompts. The probe predicts whether the prompt is in language X rather than one of the other four languages from the binary activation indicators of all SAE features at the final prompt token. We use 2,000 training prompts per language and pool activations from the shared and conflicting preference checkpoints. Features with positive probe coefficients form the target-language feature set:

$$S_{\ell,X} = \{j : \beta_j > 0\}$$

The same feature set is used when intervening on the shared and conflicting preference checkpoints.

Intervention. For a prompt in language X, we add a weighted combination of the decoder directions for features in $S_{\ell,X}$:

$$h'_\ell = h_\ell + d_\ell \, \bar{n}_{\ell,X} \, \frac{v_{\ell,X}}{\|v_{\ell,X}\|_2}, \qquad v_{\ell,X} = \sigma_\ell \sum_{j \in S_{\ell,X}} \beta_j W^{\mathrm{dec}}_{\ell,j}$$

Here, $v_{\ell,X}$ combines the selected decoder directions in proportion to their probe coefficients, and $\bar{n}_{\ell,X}$ is the mean final-token residual-stream norm at layer ℓ over 512 training prompts in language X, averaged over the shared and conflicting checkpoints. The normalized dose $d_\ell = \|h'_\ell - h_\ell\|_2 / \bar{n}_{\ell,X}$ thus sets the size of the added vector relative to a typical residual-stream norm, which makes intervention sizes comparable across models and layers. The same vector is added to both checkpoints, directly to the original residual stream and only once at the final prompt token.

We evaluate the intervention on the 100:1 shared and conflicting preference checkpoints. For each prompt, we rank the Yes/No surface forms in all five languages by mean answer-token log probability. We report English fallback, preference correctness, target-language response rate, and the rate of unparsed outputs.

Results across layers. Tables 3 and 4 report results across the evaluated layers. Under shared preferences, amplification reduces English fallback at every tested layer, with larger reductions generally appearing in later layers. Under conflicting preferences, fallback also decreases at nearly every tested layer. In both conditions, lower English fallback is accompanied by more target-language responses. Preference correctness stays close to baseline at the selected shared-preference settings and often improves under conflicting preferences. The effect is therefore not limited to a single intervention layer.

G.3 English-Associated Feature Ablation

We also intervene in the opposite direction by removing English-associated SAE features.

Feature selection. For each layer ℓ, SAE feature, and language, we fit a one-vs-rest logistic regression probe on training prompts using the feature's binary activation indicator at the final prompt token. A feature is included in the English-associated set E_ℓ if the English probe achieves the highest held-out AUC across the five languages and that AUC is at least 0.55. This selects 25–92 features per layer. The same feature set is used for the shared and conflicting preference checkpoints and for all minority languages.

Intervention. At inference time, we encode the normalized residual stream,

$$z_\ell = \mathrm{SAE}_\ell\left(\frac{h_\ell - \mu_\ell}{\sigma_\ell}\right)$$

and subtract the decoded contribution of the English-associated features:

$$h'_\ell = h_\ell - \sigma_\ell \sum_{j \in E_\ell} z_{\ell,j} W^{\mathrm{dec}}_{\ell,j}$$

All other feature contributions and the SAE reconstruction residual are left unchanged. We apply the intervention sequentially over the final five layers in one forward pass, so each later layer receives the residual stream after the preceding edits. We evaluate the intervention on all 7,176 minority-language evaluation prompts from the 100:1 checkpoints.

Results. Table 5 shows that removing English-associated features reduces English fallback for every model and preference condition. Under shared preferences, preference correctness is nearly unchanged for Llama and BLOOMZ but decreases for Qwen; Target-aligned responses nevertheless increase for all three models. Under conflicting preferences, both preference correctness and Target-aligned responses increase for all three models. The fallback reduction is smaller under conflict, where the unmodified checkpoints already have lower English fallback.

		Fallback (%)		Δ PC (pp)		Other Δ (pp)	
Layer (depth)	d_ℓ	Amplified	Δ (pp)	Raw	Balanced	Unparsed	Target lang.
Llama-3.2-1B (baseline fallback 21.7%)							
4 (0.25)	0.30	18.6	−3.1	−0.2	−0.2	−1.8	+3.2
6 (0.38)	0.30	18.9	−2.8	−0.6	−1.3	−0.2	+2.8
9 (0.56)	0.20	16.6	−5.1	+0.1	−1.2	+0.6	+5.1
10 (0.62)	0.30	9.6	−12.1	−0.8	−2.7	+2.4	+12.2
11 (0.69)	0.45	5.9	−15.8	−0.5	−2.3	+4.3	+15.8
12 (0.75)	0.20	8.5	−13.2	−0.9	0.0	+3.7	+13.1
13 (0.81)*	0.40	1.9	−19.8	−1.7	−0.8	+5.0	+19.9
14 (0.88)	0.90	0.3	−21.3	−0.1	−2.2	+2.2	+21.5
16 (1.00)	1.60	0.4	−21.3	+0.1	0.0	+0.6	+21.3
BLOOMZ-1.1B (baseline fallback 60.7%)							
6 (0.25)	0.30	46.5	−14.2	−1.6	−2.9	−0.8	+11.6
9 (0.38)	0.30	43.1	−17.6	−2.1	−2.9	−1.0	+14.7
13 (0.54)	0.30	34.7	−26.0	−1.4	−1.2	−0.9	+20.3
15 (0.62)	0.30	37.1	−23.6	−2.0	−1.8	+0.9	+19.3
17 (0.71)	0.30	23.4	−37.3	−1.2	−2.5	−0.1	+39.1
20 (0.83)*	0.60	0.0	−60.7	−0.9	+1.3	−0.9	+66.0
21 (0.88)	0.90	0.0	−60.7	−0.5	+1.1	−1.4	+66.0
22 (0.92)	0.90	0.0	−60.7	−0.4	+1.1	−1.6	+65.9
24 (1.00)	1.25	15.2	−45.5	−0.1	+1.4	−1.6	+50.7
Qwen3-1.7B (baseline fallback 62.4%)							
7 (0.25)	0.20	61.7	−0.6	−0.9	−0.5	+0.1	+0.6
11 (0.39)	0.20	54.7	−7.6	−2.5	−1.5	−0.1	+8.6
15 (0.54)	0.20	54.5	−7.9	−2.1	−1.4	−0.1	+8.5
17 (0.61)	0.10	58.5	−3.9	−1.5	−0.8	−0.1	+4.8
20 (0.71)	0.20	38.2	−24.1	−2.6	−1.4	−0.3	+26.0
22 (0.79)	0.20	37.3	−25.0	−1.8	−1.1	−0.3	+27.5
24 (0.86)*	0.60	20.3	−42.0	−0.8	−2.1	−0.3	+45.1
25 (0.89)	0.20	40.1	−22.2	−1.3	−1.1	−0.2	+24.1
28 (1.00)	0.90	18.8	−43.5	−2.2	−0.4	−0.1	+46.4
Tiny-Aya-Global-3.35B (baseline fallback 33.1%)							
10 (0.28)	0.20	29.5	−3.6	−1.3	−2.3	−0.2	+3.6
14 (0.39)	0.30	29.5	−3.6	−1.4	−1.8	−0.2	+3.6
20 (0.56)	0.20	30.9	−2.2	−0.1	−2.0	+1.0	+2.2
22 (0.61)	0.30	26.3	−6.8	−1.6	+0.4	+2.9	+6.7
25 (0.69)	0.20	25.6	−7.4	−0.9	+0.4	−1.6	+7.5
28 (0.78)	0.20	24.3	−8.7	−0.3	−0.6	−1.4	+8.8
31 (0.86)*	1.60	1.2	−31.9	−1.8	−1.5	−2.6	+32.0
32 (0.89)	1.60	3.0	−30.0	−1.9	+2.1	−2.3	+30.1
36 (1.00)	1.60	9.4	−23.7	−2.2	+1.0	−1.2	+23.8

Table 3: **Target-language feature amplification across evaluated layers: shared preferences.** For each model and layer, we report the tested intervention with the lowest English fallback subject to at most a 3 pp drop in both preference-correctness measures and at most a 5 pp increase in unparsed outputs. Fallback reports the English fallback rate after amplification and its change from the unmodified checkpoint. Raw and Balanced report changes in preference correctness; Unparsed and Target lang. report changes in the rates of unparsed and target-language outputs; PC is preference correctness. Asterisks mark the settings reported in the main paper.

		Fallback (%)		Δ PC (pp)		Other Δ (pp)	
Layer (depth)	d_ℓ	Amplified	Δ (pp)	Raw	Balanced	Unparsed	Target lang.
Llama-3.2-1B (baseline fallback 9.7%)							
4 (0.25)	0.20	9.3	−0.4	−0.1	−1.2	0.0	+0.6
6 (0.38)	0.05	9.6	−0.1	+0.2	0.0	−0.9	+0.2
9 (0.56)	0.10	9.2	−0.5	0.0	−0.6	−2.7	+0.8
10 (0.62)	0.20	7.4	−2.3	−0.4	−2.3	−1.9	+2.7
11 (0.69)	0.30	6.0	−3.7	−0.7	−0.8	−2.1	+4.1
12 (0.75)	0.40	4.9	−4.8	−0.4	−2.4	−4.4	+5.2
13 (0.81)*	1.25	0.4	−9.3	−1.4	−1.2	+2.1	+9.8
14 (0.88)	0.90	1.6	−8.1	−0.9	+0.1	−5.9	+8.5
16 (1.00)	1.25	1.8	−7.9	+0.6	+2.3	−8.1	+8.4
BLOOMZ-1.1B (baseline fallback 9.4%)							
6 (0.25)	0.30	5.5	−3.8	−2.6	+0.2	+0.7	+4.5
9 (0.38)	0.20	6.1	−3.3	−1.4	+0.4	+1.3	+3.8
13 (0.54)	0.20	4.5	−4.9	+0.3	+1.1	+0.7	+5.6
15 (0.62)	0.30	4.0	−5.4	+1.3	+1.4	+0.9	+5.7
17 (0.71)	0.45	1.4	−8.0	+0.7	−1.6	+0.4	+9.2
20 (0.83)*	0.90	0.0	−9.3	+2.6	−0.6	+0.2	+10.6
21 (0.88)	0.90	0.0	−9.3	+1.0	+2.2	−2.5	+10.6
22 (0.92)	1.25	0.1	−9.3	+4.2	+3.1	−1.3	+10.5
24 (1.00)	0.40	5.9	−3.5	+3.6	+0.4	−2.2	+4.7
Qwen3-1.7B (baseline fallback 22.6%)							
7 (0.25)	0.05	23.0	+0.4	+0.1	+0.1	+0.4	0.0
11 (0.39)	0.10	21.2	−1.4	+2.3	+2.1	−0.1	+3.9
15 (0.54)	0.20	19.8	−2.8	+3.7	+2.9	−1.2	+7.9
17 (0.61)	0.10	20.6	−2.0	+2.5	+1.4	−1.1	+5.1
20 (0.71)	0.30	13.6	−9.0	+5.0	+3.8	−4.6	+19.5
22 (0.79)	0.90	9.0	−13.6	+3.0	−1.7	−0.6	+24.7
24 (0.86)*	1.25	1.9	−20.7	+3.9	−1.4	−5.1	+31.9
25 (0.89)	1.25	3.1	−19.5	+4.0	−0.8	−3.2	+30.1
28 (1.00)	1.25	7.9	−14.7	+6.5	+6.4	+2.7	+24.0
Tiny-Aya-Global-3.35B (baseline fallback 27.8%)							
10 (0.28)	0.45	23.5	−4.3	+5.3	+5.1	−0.7	+4.6
14 (0.39)	0.60	19.1	−8.7	+6.9	+4.5	+4.2	+8.9
20 (0.56)	0.30	23.9	−4.0	+2.1	+2.6	+2.1	+4.3
22 (0.61)	0.45	12.8	−15.1	+9.2	+4.5	+4.0	+15.3
25 (0.69)	0.20	19.6	−8.3	+4.8	+1.6	−1.0	+8.6
28 (0.78)	0.20	18.5	−9.3	+3.6	+1.9	−1.0	+9.6
31 (0.86)*	1.25	3.6	−24.2	+14.2	+7.4	−1.9	+24.6
32 (0.89)	1.25	1.8	−26.0	+20.1	+9.4	−1.9	+26.4
36 (1.00)	1.25	10.3	−17.5	+10.4	+4.8	−0.7	+17.8

Table 4: **Target-language feature amplification across evaluated layers: conflicting preferences.** Columns and intervention selection follow Table 3.

	Fallback (%)			Δ PC (pp)	Δ TA (pp)
Model	Baseline	Ablated	Δ (pp)		
Shared preferences					
Llama-3.2-1B	22.6	17.3	-5.2 $[-6.7, -3.9]$	$+0.1$ $[-0.1, 0.2]$	$+4.8$
BLOOMZ-1.1B	62.8	45.5	-17.3 $[-22.0, -13.0]$	-0.4 $[-1.2, 0.4]$	$+16.0$
Qwen3-1.7B	58.2	34.9	-23.2 $[-26.2, -20.1]$	-4.4 $[-6.5, -2.2]$	$+18.0$
Conflicting preferences					
Llama-3.2-1B	9.1	7.6	-1.5 $[-1.9, -1.1]$	$+0.6$ $[0.4, 0.9]$	$+1.0$
BLOOMZ-1.1B	9.5	8.5	-1.0 $[-1.3, -0.7]$	$+0.7$ $[0.4, 1.0]$	$+0.8$
Qwen3-1.7B	24.7	20.3	-4.4 $[-5.4, -3.4]$	$+4.1$ $[3.0, 5.3]$	$+4.2$

Table 5: **Effect of ablating English-associated SAE features on English fallback.** We remove the decoded contribution of English-associated features at the final prompt token across the final five layers and evaluate the 100:1 DPO checkpoints on 7,176 minority-language prompts. Baseline and Ablated report English fallback before and after the intervention, and Δ denotes intervention minus baseline in percentage points. Brackets give 95% family-bootstrap intervals. Preference-Correct (PC) measures whether the response matches the condition-specific preferred meaning regardless of output language; Target-aligned (TA) additionally requires the response to be in the target language. Answers are ranked among four candidates (Yes and No in the target language and in English), so baselines differ slightly from Table 1.