LARA: LANGUAGE-ADAPTIVE REWARD-ANCHORED PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference optimization is typically applied to multilingual language models with one global margin scale, implicitly assuming that chosen-rejected log-probability gaps are comparable across languages. This assumption is fragile since tokenizer fertility, script, response length, and language entropy make the same semantic preference induce different numerical margins. Existing remedies reshape the loss or reweight examples, yet they still apply one global margin scale to every language. We therefore introduce Language-Adaptive Reward-Anchored Preference Optimization (LARA), which instead changes the unit of optimization while preserving the implicit-reward view of direct preference optimization (DPO). Concretely, LARA scores completions with tokenizer-independent effective-length normalization, calibrates margins with per-language scales estimated from reference gaps, and optionally adds cross-lingual anchoring and language-level group distributionally robust optimization (group-DRO). Across three backbones (Qwen3-4B, Aya Expanse 8B, and Gemma-4-12B-it) and seven languages, LARA improves average and worst-language held-out accuracy over every baseline while also balancing languages. On Gemma, where DPO barely moves the base model (51.6%/47.0%), LARA reaches 68.0%/67.3% against 60.7%/56.5% for odds-ratio preference optimization (ORPO), with no length bias and the smallest language gap of any method. Under a Friedman-Nemenyi rank test on Gemma, LARA outranks every baseline at the 5% level. Ablations and theory then isolate the mechanism, identifying effective-length normalization as the dominant driver and per-language calibration as a smaller lower-tail and balance gain, while length diagnostics rule out verbosity. The margin unit therefore decides how much accuracy a multilingual objective recovers, and the per-language scale decides how evenly that accuracy spreads across languages.

1 INTRODUCTION

Preference optimization, through reinforcement learning from human feedback and more recently direct preference optimization (DPO), is central to aligning language models with human judgment (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023). In multilingual settings these objectives use one global margin scale, treating a chosen-rejected log-probability gap as comparable across languages, a comparability that nothing in the objective enforces. Tokenizer fertility, script, response length, language entropy, and English-dominant reference priors all shift a margin's numerical scale (Rust et al., 2021; Petrov et al., 2023) without changing the semantic preference. On the tokenizer of our largest backbone, the seven languages need 0.20 to 0.63 model tokens per Unicode grapheme, and a passage of a given grapheme length therefore costs roughly three times as many tokens in Chinese as in English. Such a spread alone inflates a summed log-probability margin, and deflates a token-normalized one, before any preference evidence enters the comparison.

Data coverage alone does not explain this gap, since preference optimization already transfers across languages (Dang et al., 2024a), whereas reward benchmarks still degrade sharply outside English (Gureja et al., 2025). We therefore study the miscalibration as an objective-level failure mode. Specifically, when two languages express the same preference yet induce different reference-gap distributions, the shared temperature assigns them different gradient scales, and a model can then improve aggregate accuracy while leaving its lower tail under-aligned. Therefore, we judge a mul-

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.

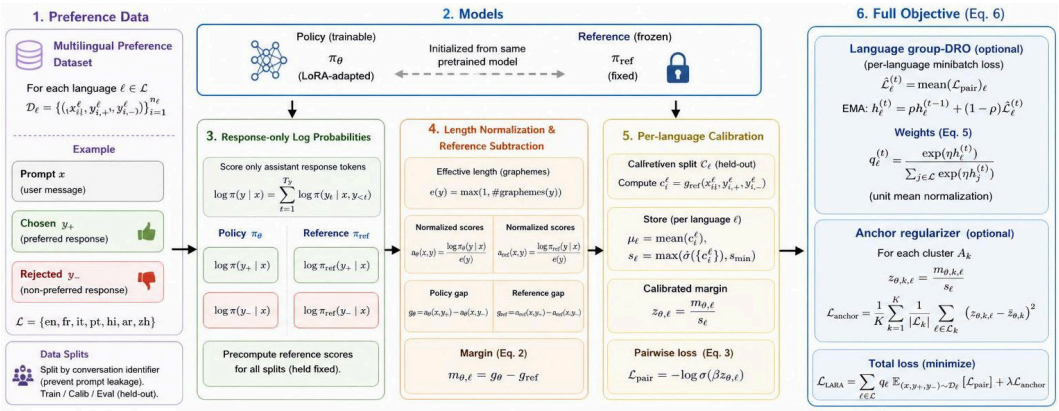


Figure 1: An overview of the complete LARA pipeline. Stages **(1)** to **(3)** depicts multilingual pairwise preferences $\mathcal{D}_\ell$, a low-rank adaptation (LoRA) policy π_θ with frozen reference π_{ref}, and response-only log-probabilities. Stage **(4)** normalizes by the grapheme-based effective length $e(y)$ and subtracts the reference, and stage **(5)** calibrates per language by s_ℓ. These two stages constitute LARA, while stage **(6)** adds the optional group-DRO and anchor terms of Sections 2.5.

tilingual objective on worst-language accuracy rather than on the average alone. By that measure DPO leaves four of seven languages at or below chance on the largest backbone we study. Its worst language falls 20.3 points below what calibrated margins reach on the same data.

We propose **LARA**, Language-Adaptive Reward-Anchored Preference Optimization, which preserves DPO's pairwise implicit-reward structure while changing the unit in which margins are compared. LARA scores completions by a tokenizer-independent effective length in Unicode graphemes, then calibrates each language's reference-subtracted margin by a scale estimated on a held-out calibration split. The scale is just one number per language, namely the dispersion of the reference model's chosen-rejected gap over at most 500 pairs per language, computed once before training without requiring any annotation. Beyond this core concept, two optional extensions add *cross-lingual anchoring*, which encourages calibrated margins of translated counterparts to agree, and *language-level group distributionally robust optimization* (group-DRO), which shifts weight toward the highest-loss language (Figure 1). Every one of these terms lives in the loss, and LARA therefore leaves the data, the model architecture, and the decoding procedure untouched.

The focus on the margin unit separates LARA from two lines of prior work. Reference-free and reshaped objectives keep one global margin scale, among them odds-ratio preference optimization (ORPO) (Hong et al., 2024), unpaired utility losses (Ethayarajh et al., 2024), bounded objectives (Azar et al., 2024), and length regularizers (Singhal et al., 2024; Park et al., 2024). Those that address length at all measure it in model tokens, a tokenizer-dependent quantity, and every one of them applies a single temperature to every language, which is exactly the choice LARA revisits. Multilingual alignment methods, by contrast, treat imbalance as a data or weighting problem that often privileges English, selecting examples by cross-lingual consistency (Zhang et al., 2025), transferring from a dominant language (Zhao et al., 2025), reweighting by confidence (Pokharel et al., 2025), or applying DRO to noisy pairs (Wu et al., 2025; Xu et al., 2025). LARA instead calibrates the margin unit itself, symmetrically across languages, and instantiates group-DRO (Sagawa et al., 2020) by treating the language as the group variable of the objective.

We evaluate LARA on Qwen3-4B (Yang et al., 2025), Aya Expanse 8B (Dang et al., 2024b), and Gemma-4-12B-it (Team et al., 2026) with seven-language LoRA preference optimization (Hu et al., 2022) on the Community Alignment dataset (Zhang et al., 2026). These languages span four scripts and a wide fertility range, which makes the margin-scale question concrete. Four research questions organize the experiments (Section 3), where a Friedman-Nemenyi rank test places LARA above every baseline at the 5% level. Our contributions are as follows:

- We formulate preference-margin miscalibration, language-dependent margin scales pooled under one temperature, as an objective-level failure mode.

- We introduce LARA, which scores responses by effective length in graphemes and calibrates margins with per-language scales, plus optional anchoring and group-DRO.
- We evaluate LARA against four baselines on three backbones and 8,514 held-out pairs per model. LARA outperforms every baseline on all three backbones and lifts Gemma average and worst-language accuracy by $+16.4$/$+20.3$ points over DPO and $+7.3$/$+10.8$ over ORPO, at the smallest language gap of any method and with no length bias.

2 METHODOLOGY

LARA keeps the DPO objective and redefines only the unit in which its chosen-rejected margin is measured. Length normalization first removes the dependence on response length and on the tokenizer, and per-language calibration then places margins from different languages on a common scale. Two optional terms, cross-lingual anchoring and language-level group-DRO, act on how margins are aggregated rather than on what they measure.

2.1 PREFERENCE SETUP

Let $\mathcal{L}$ be the set of training languages. For language $\ell \in \mathcal{L}$, the preference set is $\mathcal{D}_\ell = \{(x_i^\ell, y_{i,+}^\ell, y_{i,-}^\ell)\}_{i=1}^{n_\ell}$, where x_i^ℓ is a prompt, $y_{i,+}^\ell$ is the human-preferred completion, and $y_{i,-}^\ell$ is a rejected completion. To obtain this form, we convert multi-response examples by taking the annotated preferred response as y_+ and sampling one non-preferred non-empty response as y_-. We assign splits by conversation identifier rather than by row, and hence no prompt or conversation appears in both the training split and the held-out test split. On the model side, the trainable policy π_θ starts from the same pretrained model as the frozen reference π_{ref}, and preference scores use assistant response tokens only, $\log \pi_\theta(y \mid x) = \sum_{t=1}^{T_y} \log \pi_\theta(y_t \mid x, y_{<t})$. Against this setup, standard DPO maximizes the Bradley-Terry likelihood (Bradley & Terry, 1952) induced by the reference-regularized reward. With $\Delta_\theta = \log \pi_\theta(y_+ \mid x) - \log \pi_\theta(y_- \mid x)$ and Δ_{ref} defined analogously for the frozen reference, the DPO loss is

$$\mathcal{L}_{\text{DPO}} = -\log \text{sigmoid}\left(\beta\left[\Delta_\theta - \Delta_{\text{ref}}\right]\right). \tag{1}$$

LARA keeps this pairwise logistic form and replaces the raw log-ratio margin in Eq. 1 with a length-normalized and language-calibrated margin, because the raw margin mixes preference evidence with language-dependent nuisance variation. For efficiency, we precompute reference log-probabilities once and hold them fixed, which keeps the objective in the same implicit-reward family as DPO while avoiding a second reference pass per update.

2.2 LANGUAGE-NEUTRAL LENGTH NORMALIZATION

Summed sequence log-probabilities are sensitive to response length (Singhal et al., 2024; Park et al., 2024) and to tokenizer fertility, which varies substantially across scripts (Rust et al., 2021; Ahia et al., 2023; Petrov et al., 2023). The standard fix is dividing by the number of model tokens, is itself language-biased, because the tokens needed for the same content depend on how well the tokenizer covers each script. Over the threefold tokens-per-grapheme span (Figure 12, Appendix D), token normalization would therefore penalize high-fertility scripts far more heavily. LARA instead measures length in a unit that the tokenizer cannot distort, the effective length $e(y) = \max(1, \#\text{graphemes}(y))$. Graphemes are user-perceived characters, comparable across our Latin, Devanagari, Arabic, and Han scripts in a way that model tokens are not. We count Unicode extended grapheme clusters, which bind a base character to its combining marks and therefore treat a Devanagari syllable or a composed emoji as a single unit, and codepoints serve as the fallback wherever no segmenter is available.

The normalized completion score is $a_\theta(x, y) = \log \pi_\theta(y \mid x)/e(y)$, the policy chosen-rejected gap is $g_\theta(x, y_+, y_-) = a_\theta(x, y_+) - a_\theta(x, y_-)$, and g_{ref} is analogous under π_{ref}. LARA optimizes the reference-subtracted margin as follows:

$$m_{\theta,\ell} = g_\theta(x, y_+, y_-) - g_{\text{ref}}(x, y_+, y_-). \tag{2}$$

Similar to DPO, a positive value means that the policy has increased the normalized advantage of the preferred response over the reference. Subtracting the reference stops the loss from rewarding

responses that are already likely under the base model. This reference-anchored implicit reward motivates the name of LARA, and the translated-cluster regularizer below adds a second, cross-lingual anchor. Normalizing in this way removes a first-order length confound, since a raw DPO margin can move simply because $e(y_+) \neq e(y_-)$, even when both completions are equally likely per unit, whereas g_θ drops the length factors entirely (Proposition 1, Appendix A).

2.3 Per-Language Margin Calibration

Length normalization alone, however, does not make margins comparable across languages, because the chosen-rejected gap distribution of the reference model differs by language. LARA therefore estimates that mismatch on a held-out calibration split $\mathcal{C}_\ell$ by computing $c_i^\ell = g_{\text{ref}}(x_i^\ell, y_{i,+}^\ell, y_{i,-}^\ell)$ for each calibration pair and storing the floored calibration scale $s_\ell = \max\big(\widehat{\sigma}(\{c_i^\ell\}_{i\in\mathcal{C}_\ell}), s_{\min}\big)$, with $\widehat{\sigma}$ the sample standard deviation. The floor $s_{\min} = 0.03$ guards a language whose calibration sample carries almost no variance, which would otherwise drive β_ℓ arbitrarily high. The smallest estimated scale across the three backbones is 0.132, more than four times $s_{\min}$, and the floor therefore remains inactive. Estimating all seven scales costs one reference pass over the calibration split, which we run once before training and reuse across every method, and hence calibration leaves the per-step cost of DPO unchanged. The calibrated margin rescales Eq. 2 as $z_{\theta,\ell} = m_{\theta,\ell}/s_\ell$, and the resulting per-example loss takes the form

$$\mathcal{L}_{\text{pair}}(x, y_+, y_-, \ell) = -\log \text{sigmoid}\left(\beta z_{\theta,\ell}\right). \tag{3}$$

Eq. 3 is a DPO-style logistic objective on calibrated implicit-reward margins. Because $s_\ell > 0$, division by s_ℓ leaves the sign of every margin unchanged, and calibration therefore preserves the DPO decision boundary. The operation amounts to a language-specific temperature $\beta_\ell = \beta/s_\ell$, which scales the size of each update rather than its direction (Proposition 2). Languages with larger reference-gap variance therefore stop dominating training merely because their margins are numerically larger, while lower-variance languages still receive a meaningful share of the signal.

Calibration also makes the objective invariant to per-language nuisance scaling. When tokenizer fertility or language-specific entropy multiplies the normalized gaps of language ℓ by a factor $c_\ell > 0$, an uncalibrated DPO logit changes from $\beta m_{\theta,\ell}$ to $\beta c_\ell m_{\theta,\ell}$, whereas s_ℓ rescales by the same factor and leaves $z_{\theta,\ell}$ unchanged (Proposition 3, Appendix A). LARA therefore compares margins in standard-deviation units of the reference model, the unit that stays most stable across languages when nuisance scale rather than preference direction is the main source of mismatch.

2.4 Reward Anchoring across Translated Clusters

Calibration equalizes marginal scales without enforcing consistency across semantically matched examples. We therefore add optional anchor clusters $\mathcal{A} = \{A_k\}_{k=1}^K$, where each cluster $A_k = \{(x_k^\ell, y_{k,+}^\ell, y_{k,-}^\ell) : \ell \in \mathcal{L}_k\}$ contains translated or semantically matched versions of the same preference pair. For each member the calibrated margin $z_{\theta,k,\ell}$ and penalize the within-cluster variance is,

$$\mathcal{L}_{\text{anchor}} = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{|\mathcal{L}_k|} \sum_{\ell\in\mathcal{L}_k} \left(z_{\theta,k,\ell} - \bar{z}_{\theta,k}\right)^2, \qquad \bar{z}_{\theta,k} = \frac{1}{|\mathcal{L}_k|}\sum_{\ell\in\mathcal{L}_k} z_{\theta,k,\ell}. \tag{4}$$

The penalty in Eq. 4 falls on disagreement with the cluster mean and grants no language a privileged position in the loss (Proposition 4, Appendix A). In practice, though, all 400 clusters we build translate an English training pair outward, and English therefore remains the source language throughout (Appendix J). The term anchors calibrated preference strength rather than surface form, and the model may still answer in language-specific ways while assigning matched comparisons similar confidence.

2.5 Group-DRO over Languages

To target worst-language performance, LARA optionally applies group-DRO (Sagawa et al., 2020) with languages as groups. Let $\widehat{\mathcal{L}}_\ell^{(t)}$ be the mean of Eq. 3 over the language-ℓ examples of micro-batch t. We maintain an exponential moving average (EMA) $h_\ell^{(t)} = \rho h_\ell^{(t-1)} + (1-\rho)\widehat{\mathcal{L}}_\ell^{(t)}$ and set

the language weight as follows:

$$q_\ell^{(t)} = \frac{\exp(\eta h_\ell^{(t)})}{\sum_{j \in \mathcal{L}} \exp(\eta h_j^{(t)})}. \tag{5}$$

This softmax weighting allows smooth approximation by minimizing the maximum language loss (Proposition 5). We normalize the example weights within each micro-batch to unit mean, which keeps the loss scale comparable to non-DRO training. Because this normalization runs within a micro-batch, the reweighting acts only across examples that share one. At the micro-batch size of one used for Gemma (Appendix I) the weights reduce to one exactly, and at the size of two used for Qwen and Aya at most two examples are ever reweighted against each other. Every LARA+DRO number should therefore be read as a weak form of group-DRO rather than a test of the mechanism at full strength (Appendix J).

2.6 Full Objective and Optimization

Combining the calibrated pairwise loss, group-DRO weighting, and the anchor term yields the full objective

$$\mathcal{L}_{\text{LARA}} = \mathbb{E}_{(x,y_+,y_-,\ell)\sim\mathcal{D}} \left[\tilde{q}_\ell \, \mathcal{L}_{\text{pair}}(x, y_+, y_-, \ell)\right] + \lambda \mathcal{L}_{\text{anchor}}, \qquad \tilde{q}_\ell = \frac{q_\ell}{\bar{q}}, \tag{6}$$

where $\mathcal{D}$ is the pooled training set and $\bar{q}$ is the mean of q_ℓ over the examples of the current micro-batch. LARA proper uses $\tilde{q}_\ell = 1$ and $\lambda = 0$, which reduces the objective to the empirical mean of the calibrated pairwise loss over the pooled training set, where each language enters in proportion to its number of training pairs (Appendix I.2). Switching the optional terms on defines the extensions studied in the ablations, namely LARA+DRO with softmax q_ℓ, LARA+anchor with $\lambda > 0$, and LARA+DRO+anchor with both, while $s_\ell = 1$ defines the LARA w/o calib. ablation.

We train LoRA adapters on a frozen base model. All methods share prompt formatting, sequence length, data split, budget, optimizer, and schedule, and the LARA variants therefore differ only in which terms of Eq. 6 remain active. Learning rate and seed differ between the two families, and on Gemma the retrained ORPO run also differs in weight precision (Appendix I.1).

3 Experiments

Research Questions. Four questions organize the experiments.

- **RQ1 (accuracy and tail robustness).** Can multilingual preference optimization improve both average and worst-language held-out preference accuracy over vanilla DPO and reference-free preference optimization, holding the model, the data, and the scoring rule fixed?
- **RQ2 (mechanism).** Which objective-design choices explain any gains, whether effective-length normalization, per-language calibration, group-DRO, or the translated anchor term?
- **RQ3 (language balance and cross-lingual consistency).** Does per-language calibration reduce the language gap and produce more stable preference behavior across translated semantic clusters?
- **RQ4 (diagnostic validity).** Are the preference gains robust to length and tokenization artifacts, and do they avoid obvious free-form generation failures?

Experimental Setup. We evaluate Qwen3-4B (Qwen) (Yang et al., 2025), Aya Expanse 8B (Aya) (Dang et al., 2024b), and Gemma-4-12B-it (Gemma) (Team et al., 2026) on English (EN), French (FR), Italian (IT), Portuguese (PT), Hindi (HI), Arabic (AR), and Chinese (ZH), the last two machine-translated extensions of Community Alignment. On each we compare nine methods on the 8,514 held-out pairs, namely base, SFT, DPO, ORPO, and five LARA configurations (Section 2). Every method is scored by the normalized completion score a_θ of Section 2.2, counting a pair correct when $a_\theta(x, y_+) > a_\theta(x, y_-)$, and we report average, macro, and worst-language accuracy with the English minus non-English gap. On Gemma we retrained ORPO in bfloat16 (bf16) after a not-a-number collapse, and the retrained checkpoint was not re-run on the anchor clusters or the generation study, so its entries in those two tables are blank. Anchor clusters answer RQ3 and a

Table 1: Held-out accuracy (%) on 8,514 pairs per model, with Gap the EN minus mean non-EN accuracy. Purple is best per column, underline second-best.

	Qwen3-4B				Aya Expanse 8B				Gemma-4-12B-it			
Method	Avg	Macro	Worst	Gap	Avg	Macro	Worst	Gap	Avg	Macro	Worst	Gap
Non-LARA baselines												
Base	57.8	58.4	51.8	-4.4	58.9	59.3	51.1	-4.3	51.6	51.0	46.9	-2.5
SFT	58.2	58.7	54.0	-4.1	60.1	60.5	54.1	-4.4	54.5	54.1	51.1	-2.5
DPO	59.3	59.8	53.3	-3.4	59.9	60.3	52.1	-3.7	51.6	51.0	47.0	-2.4
ORPO	58.3	58.7	54.4	-4.1	59.7	60.1	54.4	-3.6	60.7	61.1	56.5	-2.0
LARA variants												
LARA w/o calib.	**67.6**	**67.8**	**66.3**	-1.6	65.7	65.9	63.5	-2.8	66.6	66.6	64.1	-0.2
LARA	67.2	67.5	64.9	**-1.1**	**67.0**	**67.2**	**65.4**	-1.7	**68.0**	**68.1**	**67.3**	**-0.0**
LARA+DRO	67.2	67.4	65.9	**-1.1**	**67.0**	67.1	65.3	-1.2	67.9	68.0	67.1	0.1
LARA+anchor	65.9	66.4	61.9	-2.2	66.1	66.2	63.7	**-0.1**	65.4	65.6	63.9	-0.9
LARA+DRO+anchor	65.8	66.3	61.1	-2.4	66.1	66.3	63.9	-0.2	65.2	65.4	63.8	-0.8

56-prompt generation study RQ4. Uncertainty comes from a grouped bootstrap (Efron & Tibshirani, 1994) over 5,514 conversation groups.

3.1 Held-Out Preference Accuracy

Table 1 answers the first research question (RQ1). Base, SFT, DPO, and ORPO, which compare summed log-probabilities, remain between 51.6% and 60.7% average accuracy, whereas the length-normalized LARA variants reach 65.2% to 68.0%. The ablation of Section 3.2, where length normalization alone lifts DPO by $+5.8$ points on Aya and $+15.0$ on Gemma, attributes most of that spread to the margin unit. Within the LARA family, the calibrated objective leads on Aya and Gemma, at 67.0/65.4% and 68.0/67.3%, while normalization alone leads on Qwen at 67.6/66.3%. Gemma is the starkest case, where DPO barely improves on the base model (51.6% average) and the retrained ORPO reaches only 60.7/56.5%, whereas LARA reaches 68.0/67.3%, a $+16.4$/$+20.3$ gain over DPO and $+7.3$/$+10.8$ over ORPO. Across the three backbones the gain therefore tracks the weakness of the DPO signal, rising from $+7.1$ on Aya and $+7.9$ on Qwen to $+16.4$ on Gemma. Beyond accuracy, LARA holds a -0.0 language gap and no length bias (Section 3.4). Statistically, a Friedman test on Gemma (Demšar, 2006) yields $\chi^2 = 53.5$ at $p < 0.001$. LARA+DRO and LARA obtain the best mean ranks (1.43 and 1.57), which places both more than the critical difference (CD $= 4.54$) below ORPO (6.14), SFT (7.14), DPO (8.07), and base (8.64). The other three LARA variants also outrank every baseline, yet stay within the critical difference of ORPO (Appendix C).

Beneath these aggregates the gains concentrate on the weak languages. Arabic, the worst DPO language on both smaller backbones at 53.3% and 52.1%, rises to 66.0% and 65.4% under LARA. On Gemma, where DPO leaves four of seven languages at or below chance, LARA pulls all seven into the 67 to 69% band and raises the worst from 47.0% to 67.3%. In Figure 2 the LARA intervals lie clear of the DPO and base intervals, and the gains therefore persist under grouped-bootstrap uncertainty (Appendix B). Macro and pooled accuracy also stay within 0.5 points of each other for every LARA variant, and hence the unequal per-language test sizes do not drive the aggregate.

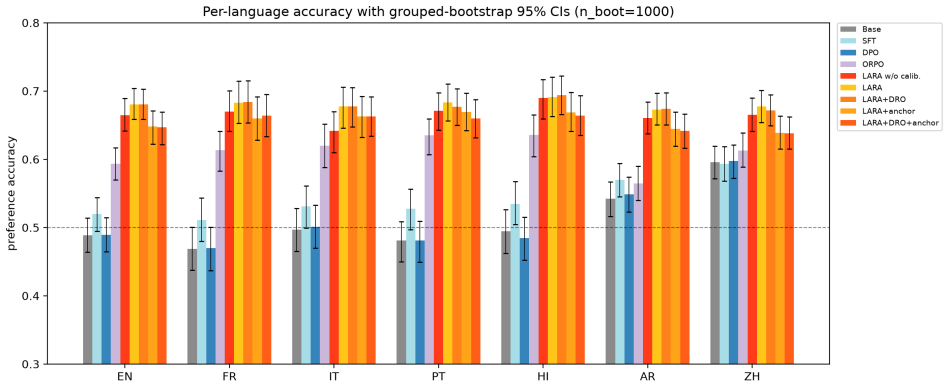


Figure 2: Per-language held-out preference accuracy on Gemma-4-12B-it, with 95% grouped-bootstrap confidence intervals.

3.2 Ablations and Statistical Uncertainty

Table 2 answers RQ2 by building LARA one component at a time on Aya, whose ladder is monotone. There DPO gains $+1.0$ average point over the untuned base, effective-length normalization adds $+5.8$, and per-language calibration adds $+1.3$ average and $+1.9$ worst-language points. Gemma follows the same order with larger steps, where normalization is worth $+15.0$ average points and lifts worst-language accuracy from 47.0% to 64.1%, and calibration adds $+1.4$ average and $+3.2$ worst-language points. Calibration is also the only rung that does not always help, since on Qwen normalization is worth $+8.3$ while the calibration step costs -0.4 average and -1.4 worst-language points, moving the gap from -1.6 to -1.1. The change of unit therefore dominates on every backbone, whereas calibration acts mainly on the tail, which is what Proposition 2 predicts for a step that rescales the gradient without moving the decision boundary. The optional extensions do not continue the ladder, since LARA+DRO matches LARA and the anchor term costs LARA 1.7 to 3.4 worst-language points across the three backbones, and LARA+DRO as much as 4.8. RQ2 therefore resolves to effective-length normalization first and calibration second, with anchoring and group-DRO shaping balance rather than accuracy. Group-DRO remains weakly tested, since the micro-batch sizes of Section 2.5 leave the reweighting nearly inactive.

Table 3 puts uncertainty on those steps. On Gemma, the median worst-language gain of LARA over DPO, SFT, and ORPO across 5,000 grouped resamples is $+19.9$, $+15.8$, and $+9.7$ points, with all three intervals above zero. These medians fall just below the point estimates of Table 1 ($+20.3$, $+16.2$, and $+10.8$) because the worst language itself can change from resample to resample. Calibration is likewise significant, at a median $+2.4$ points over LARA w/o calib. with a confidence interval (CI) of $[+0.2, +5.0]$, whereas the gap changes are not. On every backbone the worst-language gain exceeds the average gain, by $+16.2$ against $+13.5$ over SFT on Gemma, which places the improvement in the lower tail rather than in the mean.

Table 2: Component ladder on Aya Expanse 8B (Δ over the row above).

Configuration	Avg	Δ
Base (untuned)	58.9	—
DPO	59.9	$+1.0$
+ length normalization	65.7	$+5.8$
+ calibration (= LARA)	**67.0**	$+1.3$

Table 3: Paired bootstrap deltas for LARA on Gemma-4-12B-it (medians and 95% intervals, 5,000 grouped resamples).

Comparison	Worst Δ	Gap Δ
LARA – DPO	+19.9 [+17.1,+23.0]	+2.4 [-0.8,+5.6]
LARA – SFT	+15.8 [+13.2,+18.7]	+2.4 [-0.6,+5.4]
LARA – ORPO	+9.7 [+6.8,+12.6]	+2.0 [-0.8,+4.8]
LARA – w/o calib.	+2.4 [+0.2,+5.0]	+0.2 [-2.0,+2.3]

3.3 Anchor and Translated-Language Diagnostics

The anchor-cluster results of Appendix F.1 (Table 7) answer the consistency part of RQ3 on Gemma, where the clusters translate English *training* prompts outward and therefore measure translation consistency, not generalization. The calibrated variants lead, with LARA at 66.5%/62.3% average and worst-language accuracy and LARA+DRO at 66.4%/63.7%, against 51.1%/48.0% for DPO and 55.0%/51.2% for SFT. The anchor-bearing variants trail LARA by about four points despite training on these clusters, which marks anchoring as a balance regularizer. The scales calibration estimates span a factor of 3.9 on this backbone, from 0.136 for English to 0.528 for Chinese, and that spread is the mismatch the clusters expose. Cross-lingual sign agreement moves with it, rising from 0.56 to 0.73 under SFT to between 0.77 and 0.86 under LARA+DRO+anchor, with the largest gains on cross-script pairs (Appendix F.3). Ten of the eleven baseline rows score lower on the clusters than on the held-out set, by up to 4.3 points, and hence training on the English sources does not by itself carry to their translations. The pattern also holds away from Gemma, since LARA+DRO improves on DPO by $+10.3$ average and $+11.5$ worst-language points on Qwen and by $+11.7$ and $+13.0$ on Aya. Balance is the one place the clusters read less cleanly, because on Qwen every LARA variant apart from the two anchor-bearing ones still favors English by $+2.5$ to $+3.0$, which follows from building every cluster outward from an English pair.

On the held-out set every LARA variant likewise shrinks the gap relative to DPO (-2.4) and ORPO (-2.0), with LARA reaching -0.0. Calibration therefore delivers the balance that accuracy alone leaves open, closing RQ3 (Appendices F.1 and F).

Table 4: Qualitative decisions on Gemma-4-12B-it held-out pairs, where ✓ marks the human-preferred response ranked first.

Held-out pair (human preference)	Base	SFT	DPO	ORPO	LARA	LARA+DRO +anchor
FR "Peux-tu me suggérer une excursion d'une journée à Pompéi ?", where the concise itinerary (1,561 chars) is preferred over a padded variant (1,807 chars)	×	×	×	×	✓	✓
ZH (transl.) "Can you recommend hot-air balloon rides over the Champagne vineyards?", where the direct recommendation with operators and booking tips is preferred over an unrequested environmental lecture	×	×	×	×	✓	✓
FR "Quelles sont les stratégies pour gérer efficacement mon temps ?", where the thorough, structured answer (1,917 chars) is preferred over a thin single-idea reply (665 chars)	×	×	×	✓	✓	✓
AR (transl.) "What are the best beaches in Puglia?", where the answer that names beaches is preferred over a deflection into sustainability caveats	×	✓	×	×	✓	✓

3.4 Length and Normalization Robustness

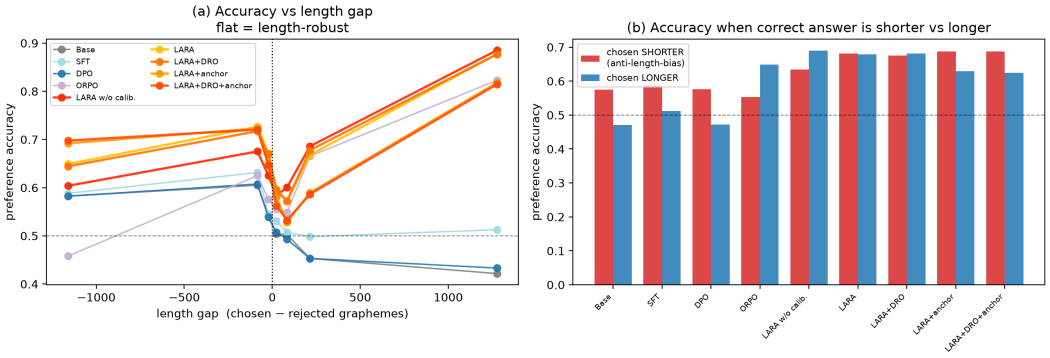


Figure 3: Length-bias diagnostic on Gemma-4-12B-it. Accuracy by chosen-rejected length gap (left) and by whether the preferred response is shorter or longer (right).

Figure 3 answers the first half of RQ4 by asking whether the gains are artifacts of length or of the tokenizer. Since the longer answer is genuinely preferred in about 57% of non-tied comparisons, a method can score above chance on length alone, and the figure therefore separates the two regimes. Within them, Gemma base and DPO stay near chance and fall slightly *lower* when the preferred answer is longer, length-bias gaps of -0.11 and -0.10. ORPO runs the other way, 0.55 when the preferred answer is shorter to 0.65 when longer, a $+0.10$ gap reflecting verbosity exploitation. LARA, by contrast, is both accurate and length-robust, near 0.68 in both regimes for a gap of 0.00, against $+0.05$ for LARA w/o calib. Resolving the same axis into deciles of the log length ratio sharpens the contrast (Appendix D.1). ORPO climbs from 0.42 in the most anti-length decile to 0.84 in the most length-aligned one and DPO falls from 0.60 to 0.44, whereas LARA moves only from 0.67 to 0.79. Length therefore explains a large share of the baseline accuracy and very little of what LARA achieves. A length-only rule would score about 57%, above base, SFT, and DPO yet below every LARA variant, which bounds length's share of the gain.

Because the held-out metric is grapheme-normalized, we also re-score every method under token normalization (Figure 13). Methods trained on a normalized margin move by at most 1.5 points, whereas the unnormalized ones move by up to 7.1 (Appendix D.3). The shift therefore reflects how a method was trained rather than calibration, and remains large enough to matter, since LARA then leads DPO by 9.3 points rather than 16.4. The two groups nonetheless keep their order, which settles the length-and-tokenization half of RQ4. Finally, Appendix G addresses the free-form half with a 56-prompt generation study that detects no verbosity inflation below the 256-token decoding cap, a screen rather than decisive evidence because the cap binds for most prompts.

3.5 Qualitative and Error Analysis

Table 4 revisits RQ1 and RQ4 on four held-out pairs. Raw sequence likelihood misleads base and DPO in both directions, ranking the padded answer first in row one and the thin answer first in row three. LARA's calibrated margins, by contrast, track the content preference both ways and recover non-Latin-script pairs that base and DPO miss (rows two and four). Rows one and three

therefore reproduce the length trap of Section 3.4 on single decisions, whereas rows two and four show the RQ1 lower tail one pair at a time. For balance, ORPO ranks the third pair correctly and SFT the fourth, consistent with the head-to-head win shares in Appendix C. The margins behind these decisions carry the same signal. On the three length traps of Appendix H.1 every baseline lands within 0.14 of the decision boundary, whereas LARA reaches $+4.10$ or more on three of the four non-English rescues. LARA's two mistakes carry margins of only -0.22 and -1.00, which keeps its errors near the decision boundary (Appendix H.4). Those residual errors spread evenly, since the LARA per-language error rate ranges only from 30.9% to 32.7%, and hence no single language accounts for what remains. Appendix H sets out eleven further pairs grouped by failure mode, including the two mistakes above with their full excerpts.

4 Related Work

Preference objectives and robust variants. Direct objectives replaced reward modeling with reinforcement learning (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022), turning the Bradley-Terry likelihood (Bradley & Terry, 1952) into a reference-subtracted log-probability ratio (Rafailov et al., 2023). Variants drop the reference (Hong et al., 2024), move to unpaired utilities (Ethayarajh et al., 2024), bound the objective (Azar et al., 2024), or penalize length (Park et al., 2024; Singhal et al., 2024). Those that address length measure it in model tokens, the very tokenizer-dependent quantity whose cross-language spread motivates this paper (Section 2.2), and all retain a global temperature that leaves the language gap LARA removes. Others reweight rather than reshape the loss, as in group-DRO (Sagawa et al., 2020) for annotation noise (Wu et al., 2025) and shift (Xu et al., 2025). Every one keeps a single margin scale, whereas LARA makes the scale per-language and the DRO groups linguistic.

Multilingual alignment and cross-lingual scale. Preference tuning transfers across languages and scales with multilingual feedback (Dang et al., 2024a), as released models (Dang et al., 2024b) and corpora (Zhang et al., 2026) show, yet reward benchmarks degrade outside English (Gureja et al., 2025). Recent remedies act on the data through consistency filtering (Zhang et al., 2025), dominant-language transfer (Zhao et al., 2025), and confidence reweighting (Pokharel et al., 2025), all of which choose which examples enter the loss, usually from English. A separate literature explains why the loss is not language-neutral, since English-dominant tokenizers fragment other scripts aggressively (Rust et al., 2021), inflating length and hence cost and context (Ahia et al., 2023; Petrov et al., 2023). Where that literature identifies a cost problem, we identify margin-scale miscalibration, which grapheme scoring removes while leaving the training data untouched.

5 Conclusion

Multilingual preference optimization should treat the margin scale as part of the alignment problem rather than as an incidental property of model likelihoods. LARA keeps the pairwise implicit-reward structure of DPO and changes only the comparison unit, scoring responses by a tokenizer-independent effective length and calibrating margins per language. Three findings support this view. *First*, the LARA variants separate from every baseline on all three backbones, and LARA itself holds the smallest language gap of any method on Gemma. *Second*, the ablations locate the mechanism, since effective-length normalization supplies most of the accuracy gain and per-language calibration adds balance and lower-tail accuracy, while anchoring shapes balance and group-DRO stays weakly tested. *Third*, LARA carries no length bias on the preference evaluation (Section 3.4), which rules out verbosity there, while the free-form check remains a screen below the decoding cap. Taken together, the margin unit decides how much accuracy a multilingual objective recovers, and the per-language scale decides how evenly that accuracy spreads. Adoption is correspondingly cheap, since the scales are seven numbers estimated once from at most 500 reference-scored pairs per language and require no annotation, no architecture change, and no additional pass over the training data.

Limitations. Our scope is limited to three open models, seven languages of which two are machine-translated, a small generation study, and family-level confounds in learning rate and seed. In the future we plan to explore and validate on more languages and multilingual models. We discuss further limitations in the Appendix J.

REFERENCES

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9904–9923, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.614. URL https://aclanthology.org/2023.emnlp-main.614/.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International conference on artificial intelligence and statistics*, pp. 4447–4455. PMLR, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4299–4307, 2017. URL https://papers.nips.cc/paper/7017-deep-reinforcement-learning-from-human-preferences.

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 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.729. URL https://aclanthology.org/2024.emnlp-main.729/.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*, 2024b.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(8):1–30, 2006. URL http://www.jmlr.org/papers/v7/demsar06a.html.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 10088–10115, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924, 1998. doi: 10.1162/089976698300017197. URL https://direct.mit.edu/neco/article/10/7/1895/6224/Approximate-Statistical-Tests-for-Comparing.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, New York, 1994. ISBN 9780412042317.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12634–12651. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ethayarajh24a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Srishti Gureja, Lester James V. Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. M-RewardBench: Evaluating reward models in multilingual settings. 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. 43–58, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.3. URL https://aclanthology.org/2025.acl-long.3/.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.626. URL https://aclanthology.org/2024.emnlp-main.626/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.297. URL https://aclanthology.org/2024.findings-acl.297/.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, volume 36, pp. 59039–59052, 2023.

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/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3118–3135, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.243. URL https://aclanthology.org/2021.acl-long.243/.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in rlhf. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=G8LaO1P0xv.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021, 2020.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.2307/2276774.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CbfsKHiWEn.

Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Robust llm alignment via distributionally robust direct preference optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

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*, 2025. URL https://arxiv.org/abs/2505.09388.

Lily H Zhang, Smitha Milli, Karen Long Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kristopher Rose, and Maximilian Nickel. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=4NtoAVqfhA.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. CM-align: Consistency-based multilingual alignment for large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 25689–25702, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1401. URL https://aclanthology.org/2025.findings-emnlp.1401/.

Weixiang Zhao, Yulin Hu, Yang Deng, Tongtong Wu, Wenxuan Zhang, Jiahe Guo, An Zhang, Yanyan Zhao, Bing Qin, Tat-Seng Chua, and Ting Liu. MPO: Multilingual safety alignment via reward gap optimization. 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. 23564–23587, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1149. URL https://aclanthology.org/2025.acl-long.1149/.

SUPPLEMENTARY MATERIAL

A Theoretical Properties of LARA
- A.1 Length Normalization and the Length Confound
- A.2 Calibration and Scale Invariance
- A.3 Anchor Agreement and Group-DRO Weighting

B Complete Held-Out Preference Results
- B.1 Cross-Model Consistency

C Rankings, Head-to-Head, and Distributional Comparisons
- C.1 Rank-Based Significance
- C.2 Worst-Case Robustness Frontier
- C.3 Paired Head-to-Head Dominance
- C.4 Per-Example Complementarity
- C.5 Grouped-Bootstrap Posteriors
- C.6 Fairness-Quality Plane

D Length and Tokenization Diagnostics
- D.1 Length Exploitation at Decile Resolution
- D.2 Tokenizer Fertility across Scripts
- D.3 Robustness to the Length Unit

E Margin Calibration and Confidence
- E.1 Reliability and Expected Calibration Error
- E.2 Calibration Scales versus Margin Dispersion
- E.3 Selective Prediction

F Cross-Lingual Consistency and Language Balance
- F.1 Complete Anchor-Cluster Results
- F.2 Held-Out Language Gap
- F.3 Cross-Lingual Transfer by Language Pair

G Generation Length Study

H Qualitative and Error Analysis
- H.1 Length Traps, Where the Human Prefers the Shorter Response
- H.2 Substance over Brevity, Where the Human Prefers the Longer Response
- H.3 Non-Latin-Script Rescues (Arabic, Hindi, Chinese)
- H.4 Error Analysis

I Implementation Details and Hyperparameters
- I.1 Models and Adaptation
- I.2 Data and Splits
- I.3 Evaluation Protocol
- I.4 Pipeline

J Limitations

K Statements
- K.1 Reproducibility Statement
- K.2 Ethics Statement
- K.3 AI Use Statement

L License

All appendix accuracies are percentages, and all tables use the same length-normalized policy-score rule as the main paper unless stated otherwise. As a guide, Table 5 maps each research question to its main-text result and to the supplementary evidence below. The main text keeps one result per research question, namely the aggregate table, the component ladder, the anchor table, and the length-bias diagnostic, together with the per-language confidence intervals and the normalization-robustness check, and every remaining figure appears here. Unless noted otherwise, moreover, figures and numbers are for the largest backbone, Gemma-4-12B-it, which we treat as the representative model. Throughout, the discussion reports absolute percentage-point deltas alongside ranked

table entries, since the central claim concerns lower-tail and cross-language robustness rather than only which method has the largest average accuracy over all pairs.

Table 5: Roadmap from research questions to main-text results and supplementary evidence.

Research question	Main-text result	Supplementary evidence
RQ1 Accuracy and tail robustness	Table 1, Figure 2, Table 3	B complete per-language tables and scaling, C rank test, dominance, bootstrap posteriors
RQ2 Mechanism	Table 2, Table 3	A proofs, D length and tokenization, E calibration and confidence
RQ3 Balance and consistency	Section 3.3	F complete anchor tables, language gap, and transfer
RQ4 Diagnostic validity	Figure 3, Table 4	G generation study, H qualitative and error analysis

A Theoretical Properties of LARA

These statements are algebraic properties of the objective rather than generalization guarantees for a nonconvex neural training problem. Within that scope, the propositions establish that the proposed normalization, calibration, anchoring, and group-DRO terms have the invariance and robustness behavior claimed in the methodology. Throughout this section, $e(y) > 0$ is the effective response length, $s_\ell > 0$ is the floored calibration scale, $\beta > 0$ is the pairwise temperature, and $\eta \geq 0$ is the group-DRO temperature. The results hold for fixed policy and reference scores, as statements about the objective itself rather than about the dynamics of training.

A.1 Length Normalization and the Length Confound

Proposition 1 (Length normalization removes a first-order length confound)**.** *Let $u_\theta(x, y) = -\log \pi_\theta(y \mid x)/e(y)$ denote the negative log-likelihood per effective unit. For a fixed preference pair, the raw policy gap satisfies*

$$\Delta_\theta = -e(y_+)u_\theta(x, y_+) + e(y_-)u_\theta(x, y_-),$$

whereas the LARA normalized policy gap is

$$g_\theta = -u_\theta(x, y_+) + u_\theta(x, y_-).$$

Holding the per-unit likelihoods fixed, the two gaps depend on the lengths in very different ways,

$$\left(\tfrac{\partial \Delta_\theta}{\partial e(y_+)}, \tfrac{\partial \Delta_\theta}{\partial e(y_-)}\right) = \big(-u_\theta(x, y_+),\, u_\theta(x, y_-)\big), \qquad \left(\tfrac{\partial g_\theta}{\partial e(y_+)}, \tfrac{\partial g_\theta}{\partial e(y_-)}\right) = (0, 0).$$

The raw gap therefore responds to length at first order whenever either per-unit negative log-likelihood is nonzero, while the normalized gap does not respond at all. In particular, if the model assigns the same per-effective-unit likelihood to both completions, $u_\theta(x, y_+) = u_\theta(x, y_-)$, then $g_\theta = 0$ whatever the lengths, while $\Delta_\theta = u_\theta(x, y_+)\,(e(y_-) - e(y_+))$ is nonzero purely because the two responses have different effective lengths.

Proof. The identities follow by substituting $\log \pi_\theta(y \mid x) = -e(y)u_\theta(x, y)$ into the definitions of the raw sequence gap and the normalized gap. The displayed derivatives are then immediate, since Δ_θ is affine in the two lengths while g_θ does not contain them. In particular, under equal per-unit likelihoods the two terms in g_θ cancel exactly, whereas the raw gap still contains $e(y_-) - e(y_+)$ and therefore assigns a nonzero margin to a pair carrying no per-unit preference evidence. The same argument applies to the reference gap, and the reference-subtracted LARA margin therefore inherits the same length-normalized unit. □

Empirically, Proposition 1 is the objective-level counterpart of the empirical length diagnostics in Appendix D, because a method whose margin is length-normalized should exhibit a flat accuracy profile across length-ratio deciles, which is what Figure 11 reports.

A.2 Calibration and Scale Invariance

Proposition 2 (Calibration preserves the pairwise decision boundary)**.** *For any language ℓ with $s_\ell > 0$, the calibrated margin $z_{\theta,\ell} = m_{\theta,\ell}/s_\ell$ has the same sign as the uncalibrated margin $m_{\theta,\ell}$. Moreover,*

$$-\log \operatorname{sigmoid}\left(\beta z_{\theta,\ell}\right) = -\log \operatorname{sigmoid}\left(\beta_\ell m_{\theta,\ell}\right), \qquad \beta_\ell = \frac{\beta}{s_\ell}.$$

Since division by s_ℓ is strictly increasing, calibration preserves the entire ordering of margins within a language, not merely their signs. Across languages, two examples with equal calibrated margins yet different scales receive update magnitudes in the ratio given by

$$\left|\frac{\partial \mathcal{L}_{\text{pair}}}{\partial m_{\theta,a}}\right| \Big/ \left|\frac{\partial \mathcal{L}_{\text{pair}}}{\partial m_{\theta,b}}\right| = \frac{s_b}{s_a},$$

and a language whose reference gaps are more dispersed therefore receives proportionally smaller updates per unit of raw margin. Calibration therefore changes the confidence and gradient scale of the logistic loss for each language, and not which response the margin prefers.

Proof. Because $s_\ell > 0$, division by s_ℓ is an order-preserving transformation, and hence $\operatorname{sign}(z_{\theta,\ell}) = \operatorname{sign}(m_{\theta,\ell})$. Substituting $z_{\theta,\ell} = m_{\theta,\ell}/s_\ell$ into the pairwise loss yields the displayed equality with $\beta_\ell = \beta/s_\ell$. Moreover, the derivative with respect to $m_{\theta,\ell}$ is

$$\frac{\partial \mathcal{L}_{\text{pair}}}{\partial m_{\theta,\ell}} = -\frac{\beta}{s_\ell} \operatorname{sigmoid}(-\beta z_{\theta,\ell}),$$

and hence the calibration scale directly controls update magnitude while preserving the decision boundary. At equal calibrated margins the sigmoid factors coincide, and the displayed ratio reduces to $(\beta/s_a)/(\beta/s_b) = s_b/s_a$. □

Proposition 2 therefore explains the component ladder of Table 2, since at a fixed policy calibration cannot flip a decision and acts instead on the gradient scale during training and on the interpretability of margin magnitudes (Appendix E).

Proposition 3 (Calibration is invariant to positive nuisance scaling)**.** *Suppose that for every language ℓ a nuisance effect such as tokenizer fertility or entropy scale multiplies both the normalized policy-reference margin $m_{\theta,\ell}$ and the reference-gap distribution used to estimate s_ℓ by its own positive factor $c_\ell > 0$. Assume the calibration floor is inactive before and after the transformation, which means that the empirical standard deviation rather than $s_{\min}$ determines s_ℓ. Then the calibrated margin is unchanged, in the sense that*

$$\frac{c_\ell m_{\theta,\ell}}{c_\ell s_\ell} = \frac{m_{\theta,\ell}}{s_\ell}.$$

Since this holds for each language separately, the calibrated objective is invariant to the entire family $\{c_\ell\}_{\ell \in \mathcal{L}}$, even when the distortions differ across languages. By contrast, an uncalibrated DPO-style logit changes from $\beta m_{\theta,\ell}$ to $\beta c_\ell m_{\theta,\ell}$, and the languages are then reweighted by whatever the nuisance factors happen to be.

Proof. For $c_\ell > 0$, empirical standard deviations scale linearly, which means that if the reference-gap random variable X has standard deviation s_ℓ, then $c_\ell X$ has standard deviation $c_\ell s_\ell$. Therefore a common positive multiplicative distortion cancels in the ratio $m_{\theta,\ell}/s_\ell$. The uncalibrated logit, by contrast, has no denominator and therefore retains the nuisance factor c_ℓ. This proves that calibration compares margins in reference-scale units rather than raw language-dependent units. If the floor is

active, exact scale invariance is no longer expected, and the floor instead acts as a conservative lower bound that prevents a language with near-zero calibration variance from receiving unbounded gradients. □

Proposition 3 concerns a per-language rescaling of the margin, and the normalization-robustness experiment of Figure 13 therefore does not test the proposition directly, since the change of unit affects every language. The experiment nonetheless points the same way, because re-scoring under a different length unit moves every method trained on a normalized margin by at most 1.5 points and the methods trained on a summed margin by up to 7.1 (Appendix D.3). Figure 15 provides the empirical support for the modeling assumption behind the proposition, namely that the estimated scales track the margin dispersion observed in the data.

A.3 Anchor Agreement and Group-DRO Weighting

Proposition 4 (Anchor loss is minimized by calibrated-margin agreement)**.** *For an anchor cluster A_k, the within-cluster anchor term*

$$\frac{1}{|\mathcal{L}_k|}\sum_{\ell\in\mathcal{L}_k}(z_{\theta,k,\ell}-\bar{z}_{\theta,k})^2$$

is the within-cluster variance of the calibrated margins. Accordingly, the term is nonnegative and equals zero if and only if $z_{\theta,k,\ell}=\bar{z}_{\theta,k}$ for every $\ell\in\mathcal{L}_k$. The cluster mean is moreover the unique center that minimizes it, since $\sum_\ell(z_{\theta,k,\ell}-c)^2$ is strictly convex in c with second derivative $2|\mathcal{L}_k|>0$. The term is invariant to permutations of the languages in the cluster and to adding a common constant to every margin in it, and hence penalizes disagreement rather than the overall level of confidence.

Proof. The term is an average of squared real numbers and hence nonnegative. The term is zero exactly when every squared term is zero, which holds if and only if every cluster member has the same calibrated margin as the cluster mean. For the center, differentiating $\sum_\ell(z_{\theta,k,\ell}-c)^2$ in c and equating to zero yields $c=\bar{z}_{\theta,k}$, and the second derivative $2|\mathcal{L}_k|$ is positive, and that center is therefore the unique minimizer. Because the definition only sums over the set $\mathcal{L}_k$ and uses the mean over that same set, reordering the languages leaves the value unchanged, and replacing every $z_{\theta,k,\ell}$ by $z_{\theta,k,\ell}+t$ shifts $\bar{z}_{\theta,k}$ by the same t and leaves each deviation untouched. Thus the regularizer targets agreement of calibrated preference strength without privileging any particular language or any particular confidence level. □

Proposition 4 constrains margin *agreement* rather than margin correctness. The anchor-bearing variants behave accordingly, leading on the cross-lingual consistency and transfer diagnostics of Appendix F while trailing on accuracy in Table 6.

Proposition 5 (Group-DRO weights emphasize higher-loss languages)**.** *For $\eta>0$, the group-DRO weights in Eq. 5 satisfy*

$$\frac{q_a^{(t)}}{q_b^{(t)}}=\exp\left(\eta\left[h_a^{(t)}-h_b^{(t)}\right]\right).$$

Therefore, if $h_a^{(t)}>h_b^{(t)}$, then $q_a^{(t)}>q_b^{(t)}$. As $\eta\to 0$ the weights become uniform, and as $\eta\to\infty$ they concentrate on the languages with maximal moving-average loss. The weights are moreover the gradient of the smoothed maximum

$$\Phi_\eta(h)=\frac{1}{\eta}\log\sum_{\ell\in\mathcal{L}}\exp(\eta h_\ell),\qquad \frac{\partial\Phi_\eta}{\partial h_\ell}=q_\ell,$$

which brackets the worst-language loss,

$$\max_{\ell\in\mathcal{L}}h_\ell\;\le\;\Phi_\eta(h)\;\le\;\max_{\ell\in\mathcal{L}}h_\ell+\frac{\log|\mathcal{L}|}{\eta}.$$

The group-DRO weights therefore approximate gradient descent on a surrogate that never underestimates the worst-language loss and exceeds it by at most $\eta^{-1} \log |\mathcal{L}|$, which for our seven languages and $\eta = 1$ is $\log 7 \approx 1.95$ in loss units and shrinks as η grows. The correspondence is approximate because the implementation computes the weights from the moving average and renormalizes them within each micro-batch, as Section 2.5 describes.

Proof. Taking the ratio of the two softmax weights cancels the common denominator and yields the displayed expression. Since the exponential function is strictly increasing, larger moving-average loss implies larger weight. The limits, in turn, are standard softmax limits. With $\eta = 0$ all exponentials equal one, and as η grows the maximum-loss terms dominate every smaller term at an exponential rate. Differentiating Φ_η yields $\partial \Phi_\eta / \partial h_\ell = \exp(\eta h_\ell) / \sum_j \exp(\eta h_j) = q_\ell$. For the bracket, let $M = \max_\ell h_\ell$. Every term of the sum is at most $\exp(\eta M)$ and at least one equals it, and hence $\exp(\eta M) \le \sum_\ell \exp(\eta h_\ell) \le |\mathcal{L}| \exp(\eta M)$, and taking $\eta^{-1} \log$ of each side yields the two inequalities. This proves that the group-DRO component approximates worst-language weighting smoothly, and the two bounds above control the approximation error. □

Taken together, the five propositions separate the two components that change what a margin *is* (Propositions 1 to 3) from the two that change how margins are *aggregated* (Propositions 4 and 5). The ablations of Section 3.2 recover the same split empirically, where the first group supplies the accuracy gains and the second shapes language balance together with worst-language behavior.

B Complete Held-Out Preference Results

Table 6 reports the full per-language held-out breakdown for every method, expanding the evidence for RQ1 and RQ3. This table combines the per-language numbers summarized in Section 3.1 with the aggregate metrics of Table 1 in a single reference table, and covers all three backbones rather than the representative Gemma backbone of Figure 2. The complete table reinforces the main RQ1 conclusion. On Aya, no unnormalized baseline exceeds 60.1% average accuracy, and DPO remains at 59.9%. LARA and LARA+DRO, by contrast, reach 67.0% average accuracy, with LARA achieving the best worst-language score at 65.4%. The per-language columns indicate that no single language drives this gain. In particular, LARA improves the weakest DPO language, Arabic, from 52.1% to 65.4%, while also raising English accuracy from 57.1% to 65.7%.

The effect sizes are large relative to the baseline spread. Compared with DPO, the best LARA variants gain 7.1 average points and 13.3 worst-language points. LARA improves every language, gaining +8.6 on English, +7.5 on French, +5.4 on Italian, +6.7 on Portuguese, +1.7 on Hindi, +13.3 on Arabic, and +4.6 on Chinese. This pattern strengthens the RQ1 claim because the LARA gain does not come from sacrificing one language to improve another, and the largest absolute correction appears exactly in the DPO lower tail. For LARA, grouped-bootstrap estimates yield a worst-language improvement over DPO of $+12.6$ points with a 95% interval of $[+9.5, +15.6]$. Thus the lower-tail improvements are not explained by a small number of conversation groups.

The Gemma block exhibits the same structure in a more extreme regime. There DPO is statistically indistinguishable from the untuned base (51.6% average, 47.0% worst-language), and four of its seven languages sit at or below chance, including English at 48.9%, and the usual reference-anchored preference signal is therefore almost absent on this backbone. The retrained ORPO recovers a real signal (60.7% average) yet leaves a 56.5% Arabic tail and a -2.0 language gap. LARA lifts all seven languages into the 67 to 69% band and improves the weakest DPO language, French, from 47.0% to 68.3%, for a $+16.4$ average and $+20.3$ worst-language gain over DPO and $+7.3$/$+10.8$ over ORPO. Every LARA language thus lands above the best baseline language on this backbone. Resampling in turn shifts these figures slightly, because the worst language can change from resample to resample. The grouped bootstrap for LARA yields a median worst-language improvement of $+19.9$ points over DPO (95% interval $[+17.1, +23.0]$) and $+9.7$ over ORPO ($[+6.8, +12.6]$), the largest and most tightly bounded lower-tail gains in the study, with both intervals clear of zero.

Finally, this per-language view identifies the lower tail that the calibration diagnostics of Appendix E explain. French is the worst DPO language on Gemma at 47.0%, and LARA lifts it to 68.3% while

Table 6: Complete seven-language held-out preference results (8,514 test pairs per model). Gap is EN minus mean non-EN accuracy.

Method	Avg	Macro	Worst	Gap	EN	FR	IT	PT	HI	AR	ZH
Qwen3-4B											
Non-LARA baselines											
Base	57.8	58.4	51.8	-4.4	54.6	60.9	60.3	60.8	61.3	51.8	58.8
SFT	58.2	58.7	54.0	-4.1	55.2	60.6	60.2	60.3	60.9	54.0	59.5
DPO	59.3	59.8	53.3	-3.4	56.9	62.4	61.9	62.0	62.7	53.3	59.6
ORPO	58.3	58.7	54.4	-4.1	55.2	60.3	60.1	60.4	60.5	54.4	59.9
LARA variants											
LARA w/o calib.	**67.6**	**67.8**	**66.3**	-1.6	66.5	68.5	66.3	**68.3**	**72.4**	**66.3**	**66.3**
LARA	67.2	67.5	64.9	**-1.1**	**66.6**	69.1	67.1	68.1	70.9	66.0	64.9
LARA+DRO	67.2	67.4	65.9	**-1.1**	66.5	**69.3**	67.7	66.4	70.3	66.0	65.9
LARA+anchor	65.9	66.4	61.9	-2.2	64.5	68.5	**69.3**	66.8	67.8	61.9	65.8
LARA+DRO+anchor	65.8	66.3	61.1	-2.4	64.3	68.5	69.2	67.0	68.2	61.1	65.7
Aya Expanse 8B											
Non-LARA baselines											
Base	58.9	59.3	51.1	-4.3	55.7	59.3	59.2	62.9	65.1	51.1	62.1
SFT	60.1	60.5	54.1	-4.4	56.7	60.0	62.2	62.7	65.5	54.1	62.5
DPO	59.9	60.3	52.1	-3.7	57.1	60.3	60.6	62.8	66.5	52.1	63.0
ORPO	59.7	60.1	54.4	-3.6	57.0	59.6	62.1	61.9	63.5	54.4	62.0
LARA variants											
LARA w/o calib.	65.7	65.9	63.5	-2.8	63.5	65.6	65.1	68.4	68.5	64.7	65.6
LARA	**67.0**	**67.2**	**65.4**	-1.7	65.7	**67.8**	**66.0**	**69.5**	68.2	**65.4**	67.6
LARA+DRO	**67.0**	67.1	65.3	-1.2	**66.1**	**67.8**	65.8	68.9	68.3	65.3	**67.8**
LARA+anchor	66.1	66.2	63.7	**-0.1**	**66.1**	66.6	65.2	66.6	**69.2**	63.7	66.2
LARA+DRO+anchor	66.1	66.3	63.9	-0.2	**66.1**	66.6	65.5	66.6	69.0	63.9	66.2
Gemma-4-12B-it											
Non-LARA baselines											
Base	51.6	51.0	46.9	-2.5	48.9	46.9	49.7	48.1	49.4	54.2	59.6
SFT	54.5	54.1	51.1	-2.5	52.0	51.1	53.1	52.8	53.5	57.0	59.3
DPO	51.6	51.0	47.0	-2.4	48.9	47.0	50.1	48.1	48.4	54.9	59.7
ORPO	60.7	61.1	56.5	-2.0	59.3	61.3	62.0	63.5	63.5	56.5	61.3
LARA variants											
LARA w/o calib.	66.6	66.6	64.1	-0.2	66.5	67.0	64.1	67.1	69.0	66.1	66.5
LARA	**68.0**	**68.1**	**67.3**	**-0.0**	**68.1**	68.3	**67.7**	**68.3**	69.1	67.3	**67.7**
LARA+DRO	67.9	68.0	67.1	0.1	**68.1**	**68.4**	**67.7**	67.7	**69.4**	**67.4**	67.1
LARA+anchor	65.4	65.6	63.9	-0.9	64.8	66.0	66.3	66.9	66.9	64.5	63.9
LARA+DRO+anchor	65.2	65.4	63.8	-0.8	64.7	66.4	66.3	66.0	66.4	64.1	63.8

also raising English from 48.9% to 68.1%, and the correction is therefore not a redistribution of accuracy from the strong languages to the weak ones.

B.1 Cross-Model Consistency

Figure 4, together with Tables 1 and 6, confirms that the same pattern holds across all three backbones, from the 4B Qwen model up to the 12B Gemma model. The LARA family band occupies the top of the range on both macro and worst-language accuracy at every scale, and no baseline reaches into that band. Every baseline stays well below the band at all three scales. The largest per-language corrections consistently land on the weakest DPO language, Arabic on both smaller backbones, where LARA w/o calib. gains 13.0 points on Qwen. Hindi reaches the highest single per-language accuracy in the study, 72.4% under the same configuration, up from 62.7% for DPO.

Gemma is the most demanding replication and the sharpest test of the normalization principle. Its base and DPO checkpoints lie near or below chance on most languages (51.6% average, 47.0% worst-language for DPO), and SFT adds only 2.9 points, and hence neither supervised fine-tuning nor reference-anchored preference optimization contributes usable signal on this backbone. ORPO

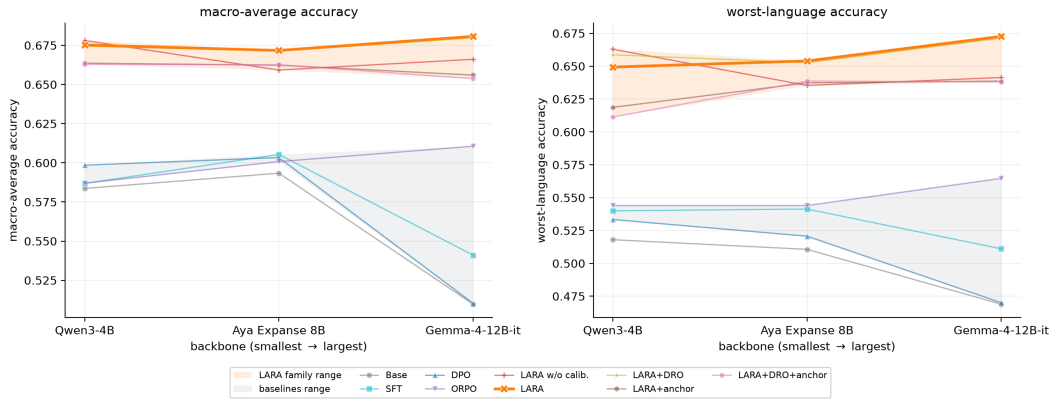


Figure 4: Scaling consistency across backbones, with bands spanning the LARA family (pink) and the baselines (gray). ORPO forms the upper gray edge.

is the only unnormalized baseline that escapes this regime (60.7% average, 56.5% worst-language), while LARA reaches 68.0% average and 67.3% worst-language accuracy, for a $+16.4/+20.3$ point gain over DPO and $+7.3/+10.8$ over ORPO. The length-normalized family therefore recovers a strong signal exactly where the unnormalized ones fail. Because LARA w/o calib. differs from DPO only in the length unit and already reaches 66.6% here, the recovery is attributable to that unit rather than to the surrounding objective. Beyond the point estimates, a grouped bootstrap ($n = 5{,}514$ conversation groups) confirms that the lower-tail gains over the unnormalized baselines are not driven by a few groups, since LARA improves worst-language accuracy by $+19.9$ points over DPO (95% CI $[+17.1, +23.0]$) and $+9.7$ points over ORPO (95% CI $[+6.8, +12.6]$). Across three architectures spanning 4B to 12B parameters, then, LARA consistently lifts the lower tail, and the size of the lift grows precisely on the backbone where uncalibrated preference optimization is weakest, and Appendix C recovers the same ordering example by example on the individual held-out pairs.

C Rankings, Head-to-Head, and Distributional Comparisons

This section reports the ranked, paired, and resampled comparisons that place the aggregate RQ1 accuracies of Table 1 on a statistical footing. All analyses use the Gemma-4-12B-it held-out test set and the common length-normalized policy-score rule. The six views answer progressively finer questions. The first asks which method ranks highest across languages, the second how the whole per-language profile compares, the third and fourth which method wins the examples that two methods disagree on, the fifth how much of that ordering persists under resampling, and the sixth how accuracy trades off against language balance.

C.1 Rank-Based Significance

Figure 5 depicts the Friedman-Nemenyi analysis reported in Section 3.1, treating each of the seven languages as an independent trial. The Friedman test over per-language accuracies (Demšar, 2006) yields $\chi^2 = 53.5$ ($p < 0.001$) over the nine methods. LARA+DRO and LARA take the best mean ranks (1.43 and 1.57), and both lie more than the critical difference (CD $= 4.54$) below ORPO (6.14), SFT (7.14), DPO (8.07), and base (8.64). Their margins over ORPO, 4.71 and 4.57 points of rank, clear the threshold by a narrow amount, and with seven languages as the only trials the test has limited power at that resolution. Between the two groups fall the w/o-calibration and anchor-bearing configurations, which anticipates the ablation ordering of Section 3.2. Overall, then, the rank test separates the calibrated LARA configurations from every baseline, and inside the LARA family it ranks those configurations first, although the remaining differences within the family stay below the critical difference of the test.

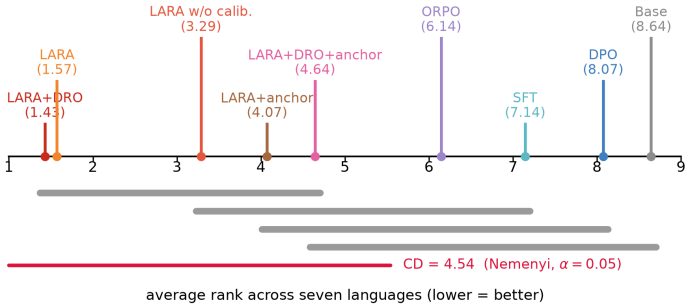


Figure 5: Critical-difference diagram (Friedman-Nemenyi, $\alpha = 0.05$) on Gemma-4-12B-it. A bar joins statistically indistinguishable methods.

C.2 Worst-Case Robustness Frontier

Figure 6 redraws the same evidence as a frontier. For each method we sort the seven per-language accuracies from worst to best, and the leftmost point is therefore the worst-language accuracy while the curve traces the whole lower tail rather than a single summary. Under that construction, a method with genuinely balanced multilingual behavior produces a high, flat curve, whereas a method that carries one weak language produces a curve that starts low and climbs. By that reading, the LARA variants lie far above base, SFT, and DPO and clear ORPO across the entire frontier, not only at the worst-language point that Table 1 reports. LARA in particular starts at 0.673 and stays between 0.673 and 0.691 across all seven languages, which makes it the flattest curve in the figure. This flatness is precisely the frontier form of the language gap of Table 1, since a method can reach a given worst case and still distribute the rest unevenly. The frontier thus restates the ordering of Section 3.2 in place of the component ladder (Table 2). The same ordering is why we report worst-language accuracy as a headline metric, since for these methods the choice of lower-tail quantile never changes which group a method belongs to.

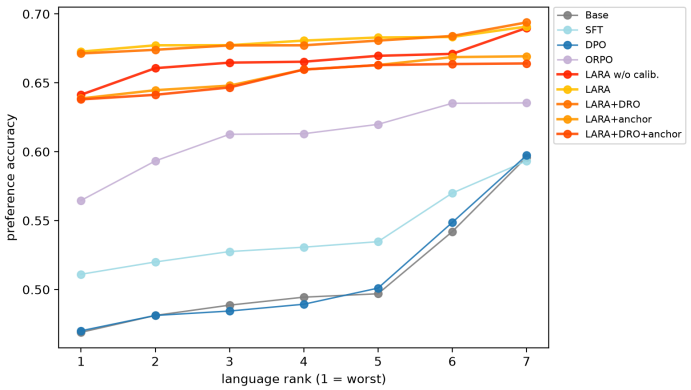


Figure 6: Robustness frontier on Gemma-4-12B-it, with per-language accuracies sorted from worst to best. Higher curves on the left indicate stronger tails.

C.3 Paired Head-to-Head Dominance

Figure 7 reports McNemar-style head-to-head comparisons (Dietterich, 1998). For each method pair, the figure reports the share of discordant examples, where one method is correct and the other wrong, won by the row method, with significance at $p < 0.05$. LARA wins 71% of its disagreements with DPO, 70% with SFT, and 62% with the retrained ORPO, all significant, and 55% with LARA w/o calib. ORPO in turn wins 62% of its disagreements with DPO, and the head-to-head ordering (LARA $\succ$ ORPO $\succ$ DPO/SFT/base) therefore matches the critical-difference ranking in Figure 5.

Because this test conditions on the examples where two methods actually disagree, it rules out the possibility that compensating errors hide behind equal aggregate accuracies.

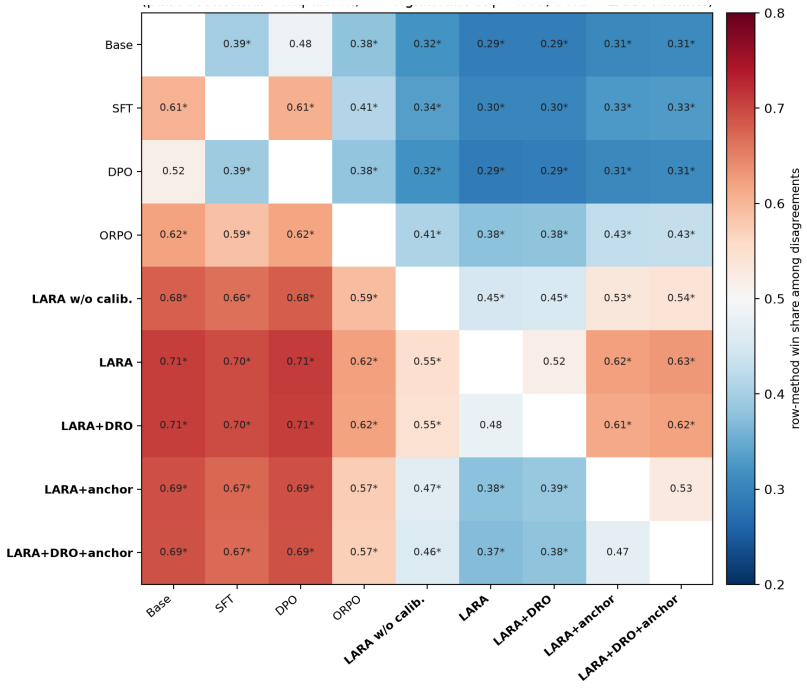


Figure 7: Head-to-head dominance on Gemma-4-12B-it. Entries are the share of discordant examples won by the row method (* marks $p < 0.05$).

C.4 Per-Example Complementarity

The disagreement heatmap (Figure 8) indicates that the gain is not a uniform shift, because LARA is correct on 28% of examples where DPO is wrong (11% in the reverse direction) and on 18% of examples where ORPO is wrong (11% in reverse). The same asymmetry holds against SFT and base, which confirms that the advantage of LARA concentrates on exactly the examples that the baselines get wrong rather than spreading evenly across the whole held-out test set. Within the LARA family the picture inverts, since LARA and LARA+DRO disagree on only 2.8% of the test set against 39% disagreement with DPO. In short, the margin unit rather than the surrounding objective decides which examples a method gets right.

C.5 Grouped-Bootstrap Posteriors

Figure 9 provides the resampled view for LARA. Across 800 grouped-bootstrap resamples its worst-language improvement is positive in every resample against DPO (median $+19.7$ points), SFT ($+15.7$ points), and ORPO ($+9.6$ points). Turning to balance, the right panel plots gap *closure*, $|\mathrm{gap}_{\mathrm{baseline}}| - |\mathrm{gap}_{\mathrm{LARA}}|$, which is positive in 82% of resamples against SFT, 81% against DPO, and 76% against ORPO. Gap closure is a different quantity from the signed gap delta of Table 3, and the two should be read separately. The medians here differ from that table in the first decimal ($+19.7$ against $+19.9$ for DPO, for instance) because this figure uses 800 resamples where the table uses 5,000. The contrast between the two panels remains informative, because the worst-language gain over the unnormalized baselines is unambiguous under resampling, whereas the gap improvement is a directional yet not significant effect, and the main text therefore states the worst-language and gap claims with different strength.

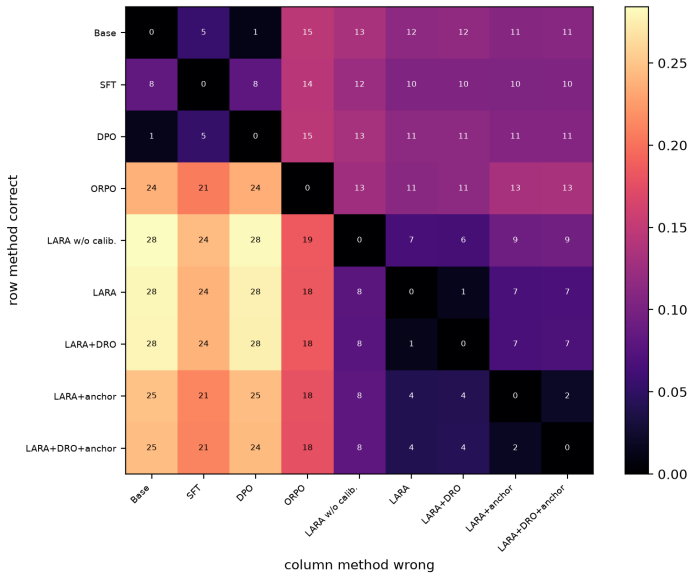


Figure 8: Pairwise disagreement diagnostic on Gemma-4-12B-it. Each entry counts cases where one method is correct and the other is wrong.

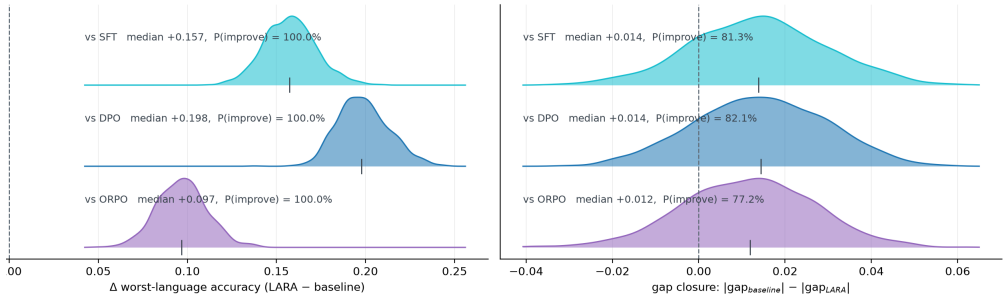


Figure 9: Bootstrap posteriors (800 resamples) of the LARA gain over each Gemma baseline in worst-language accuracy (left) and in gap closure, $|\mathrm{gap}_{\mathrm{baseline}}| - |\mathrm{gap}_{\mathrm{LARA}}|$ (right).

C.6 Fairness-Quality Plane

Figure 10 places every method on the plane of macro accuracy versus the language gap, with 2σ bootstrap ellipses. LARA and LARA+DRO sit closest to the ideal corner of high accuracy at a zero gap, and their ellipses do not overlap those of base, SFT, DPO, or ORPO. The joint accuracy-and-balance claim of RQ1 and RQ3 therefore holds under resampling uncertainty rather than resting on point estimates alone. The anchor-bearing variants, by contrast, sit lower on the accuracy axis while remaining near zero gap, and their ellipses overlap those of the calibrated variants on the balance axis alone. Reading the two axes together also rules out the cheapest way to close a language gap, namely lowering English accuracy, because the calibrated variants are simultaneously among the most accurate and the most balanced.

D Length and Tokenization Diagnostics

This section completes the RQ2 mechanism evidence on the two quantities LARA changes in the scoring rule, response length and the length unit. Figure 11 resolves the accuracy-versus-length picture of Figure 3 at decile resolution, Figure 12 quantifies the tokenizer disparity that motivates

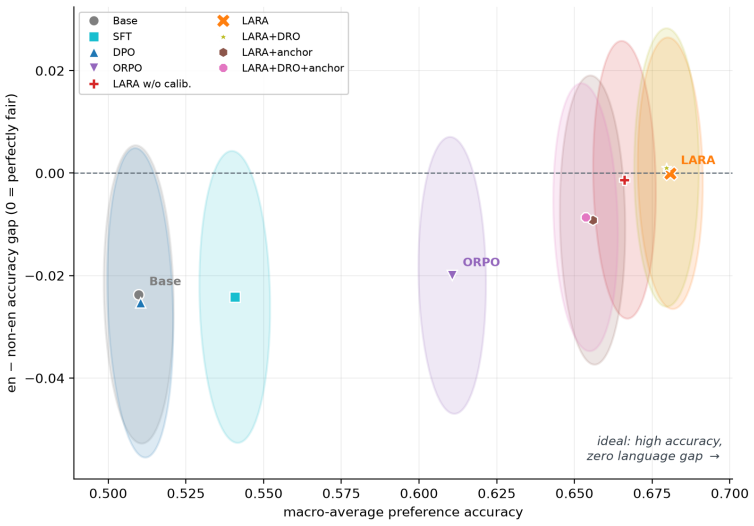


Figure 10: Fairness-quality plane on Gemma-4-12B-it with 2σ grouped-bootstrap ellipses (800 resamples). The ideal region is high macro accuracy at zero language gap.

effective-length normalization in Section 2.2, and Figure 13 re-scores every method under the alternative length unit.

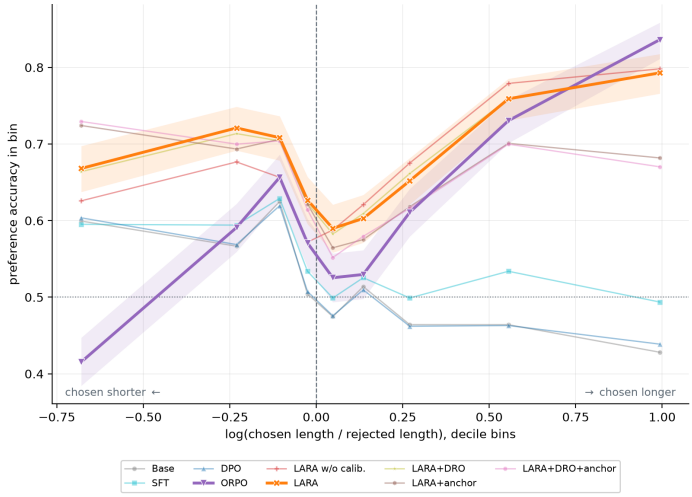


Figure 11: Accuracy across deciles of the log chosen/rejected length ratio on Gemma-4-12B-it, with 95% Wilson bands. Flat curves indicate content-based judgments.

D.1 Length Exploitation at Decile Resolution

Figure 11 stratifies accuracy by deciles of the log chosen/rejected length ratio rather than by a binary shorter/longer split. Here the shape of each curve is diagnostic, since a method judging content should stay roughly flat while one exploiting verbosity climbs to the right. By that test, the retrained ORPO has the steepest slope of any method, rising from 0.42 in the leftmost decile (preferred answer much shorter) to 0.84 in the rightmost, and its Wilson bands (Wilson, 1927) separate the two ends cleanly. The span sharpens the reading of its Gemma recovery, since ORPO is weakest exactly where the preference contradicts length. This regime is precisely the one a multilingual objective cannot afford to fail in, because the length signal is itself tokenizer-dependent and language-dependent.

The LARA curves, by contrast, stay comparatively flat and lead on the anti-length side, and their advantage is therefore not inherited from the length prior.

D.2 Tokenizer Fertility across Scripts

The tokenizer diagnostic provides the numbers behind the claim that model-token length is not a language-neutral unit. Average model tokens per grapheme vary substantially across scripts (Figure 12), with Latin-script languages (≈ 0.20) having much lower ratios than Hindi (0.38) and Chinese (0.63), a factor of three between the two extremes. Correspondingly, Figure 13 confirms that the length-normalized variants are insensitive to this disparity while the summed-margin baselines absorb it, which is why LARA measures effective length in graphemes rather than in model tokens.

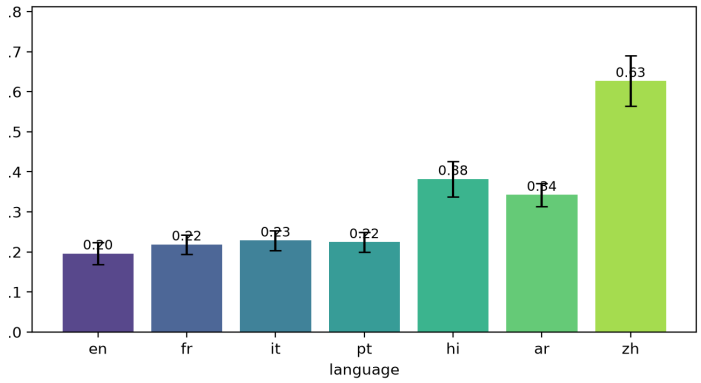


Figure 12: Model tokens per Unicode grapheme by language for the Gemma-4-12B-it tokenizer, from 0.20 (English) over 0.38 (Hindi) to 0.63 (Chinese).

D.3 Robustness to the Length Unit

Figure 13 re-scores every method under token normalization instead of graphemes. The methods trained on a length-normalized margin barely move, since LARA shifts by 0.0 points, LARA+DRO and both anchor variants by 0.4, and LARA w/o calib. by 1.5. The methods trained on a summed margin move a great deal, by 7.1 (DPO), 6.9 (base), 5.4 (SFT), and 2.4 (ORPO), always upward, because token normalization supplies at scoring time part of the correction they never learned. Stability is therefore a property of training on a normalized margin rather than of per-language calibration specifically, since LARA w/o calib. is uncalibrated and still nearly stable. Proposition 3 predicts invariance to a per-language rescaling of the margin, a narrower statement than this figure tests. The figure adds that the group ordering of Table 1 persists under a change of scoring unit, whereas the size of the gaps does not (Section 3.4).

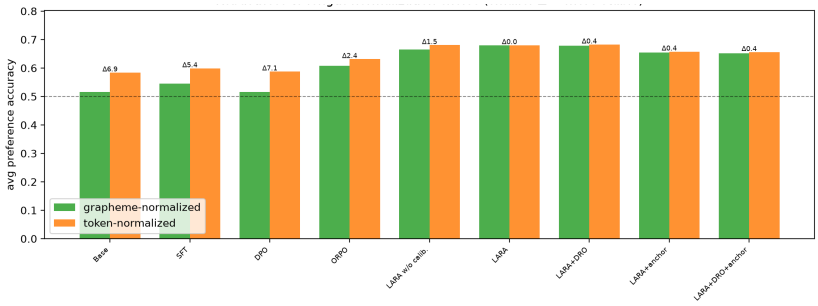


Figure 13: Normalization robustness on Gemma-4-12B-it under grapheme-normalized versus token-normalized scoring, with Δ their difference.

E MARGIN CALIBRATION AND CONFIDENCE

Calibration also changes what a margin *means*, the part of RQ2 that accuracy alone cannot reveal (Section 3.2). The three diagnostics here ask whether margin magnitude is trustworthy (Figure 14), whether the estimated language scales are data-grounded (Figure 15), and whether the resulting confidence ordering is useful in practice (Figure 16).

E.1 RELIABILITY AND EXPECTED CALIBRATION ERROR

Figure 14 asks whether a larger preference margin corresponds to a higher empirical probability of being correct. Every LARA configuration tracks the diagonal closely, whereas base and DPO deviate from it sharply. Quantitatively, the expected calibration error (ECE, Guo et al., 2017) reaches 0.123 for DPO and 0.124 for the base model, against 0.034 for LARA w/o calib., 0.041 for LARA, 0.040 for LARA+DRO, and 0.046 for ORPO. Base and DPO therefore carry roughly three times the ECE of the calibrated LARA configurations, whereas ORPO remains within a factor of 1.2 of LARA. The anchor term rather than per-language calibration improves this diagnostic further, since LARA+DRO+anchor reaches 0.024 and LARA+anchor 0.025, the two lowest values in the figure. Calibration error is the one diagnostic on which the anchor-bearing variants lead, and the outcome agrees with Proposition 4, which constrains margin agreement rather than margin correctness. Per-language calibration on its own, then, does not lower ECE relative to normalization alone, and we therefore make no calibration-of-confidence claim for the step. Its contribution is the language balance of Section 3.3. Comparable magnitudes still matter for the anchoring and group-DRO terms, since both consume the magnitude of a margin and not only its sign.

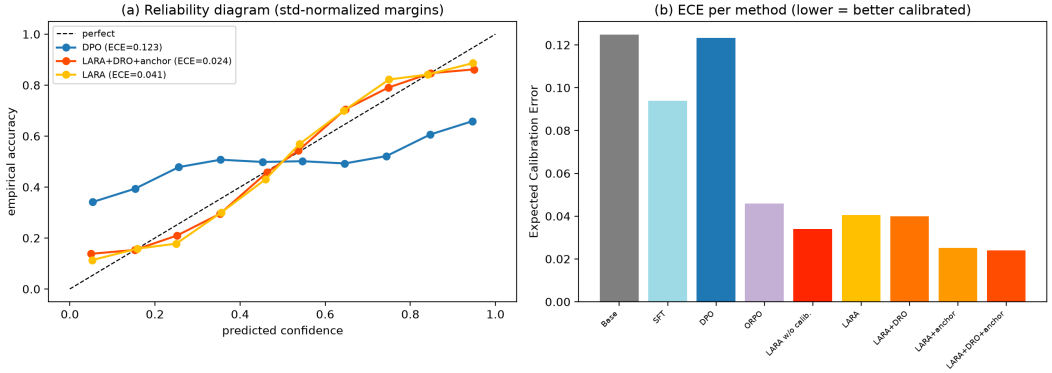


Figure 14: Calibration of preference margins on Gemma-4-12B-it. Panel (a) plots empirical accuracy against predicted confidence, where a well-calibrated method tracks the diagonal, and panel (b) reports expected calibration error per method.

E.2 CALIBRATION SCALES VERSUS MARGIN DISPERSION

The margin-versus-scale diagnostic (Figure 15) checks the central calibration assumption of LARA and the modeling premise of Proposition 3. The estimated calibration scales track empirical test-margin spread closely ($r = 0.95$), with larger scales for high-dispersion languages such as Chinese and Hindi and smaller scales for the Latin-script languages. This in turn supports treating the scale as a data-grounded unit conversion estimated on a calibration split rather than as an arbitrary per-language weight, and it is the reason the calibration floor $s_{\min}$ is rarely active in practice.

E.3 SELECTIVE PREDICTION

The selective-prediction curves (El-Yaniv & Wiener, 2010) (Figure 16) demonstrate that LARA margins are informative, not only positive more often. The LARA family occupies the top of the range on area under the accuracy-coverage curve (AUC), from 0.72 for the anchor variants to 0.75 for LARA+DRO, compared with 0.52 for DPO, 0.56 for SFT, and 0.62 for ORPO. The most confident tenth of examples is about 87% accurate for LARA and LARA+DRO, well above every baseline

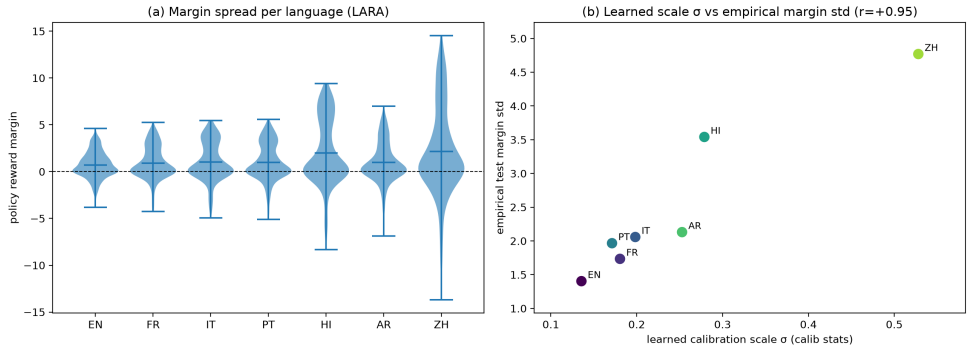


Figure 15: Calibration scales against margin dispersion on Gemma-4-12B-it. Panel (a) plots the per-language spread of LARA margins, and panel (b) plots the learned scale against the empirical margin standard deviation ($r = 0.95$).

at that coverage. A normalized margin therefore carries a practical benefit, because margins on that scale support abstention and confidence thresholding, which the summed DPO margins on this backbone do not.

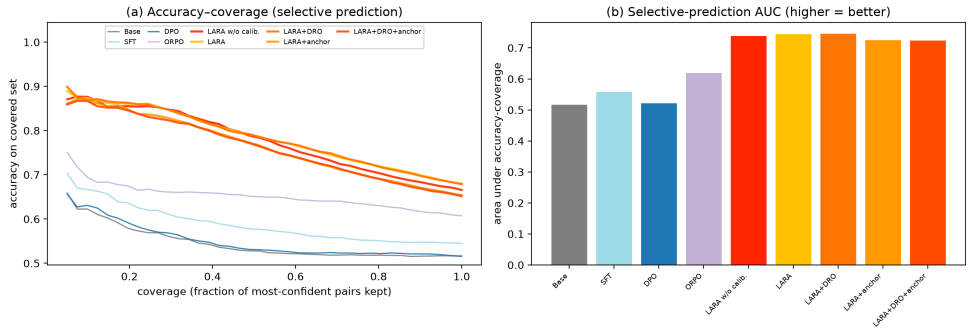


Figure 16: Selective prediction on Gemma-4-12B-it. Panel (a) traces accuracy as each method abstains on low-confidence examples, and panel (b) reports the area under that accuracy-coverage curve.

F Cross-Lingual Consistency and Language Balance

F.1 Complete Anchor-Cluster Results

Table 7 reports the complete translated anchor-cluster breakdown for RQ3, expanding the Gemma summary of Section 3.3 to all three backbones. These 2,800 examples come from 400 semantic clusters, each evaluated across the seven-language translated set. Every cluster is a machine translation of an English pair drawn from the *training* split, and the table is therefore neither a held-out generalization test nor comparable to Table 6 in that respect. The table instead isolates whether one preference decision carries over under translation, a question the held-out set leaves open because its items have no matched counterparts. Two caveats follow. First, LARA+anchor and LARA+DRO+anchor trained on exactly these clusters, and their rows therefore give the most favorable reading of the anchor term. Second, every method has seen the English source pairs during training, and the absolute numbers here therefore measure consistency under translation rather than generalization. LARA and LARA+DRO lead on Aya and Gemma, and on Qwen LARA+DRO ties LARA w/o calib. for the lead, while the anchor-bearing variants mainly improve language balance, consistent with the RQ2 finding that anchoring acts more as a regularizer than as the main accuracy driver. LARA+DRO in turn reaches 65.4%/63.7% on Qwen and 67.6%/65.5% on Aya, taking the best worst-language anchor score on Qwen and matching LARA for it on Aya.

Table 7: Complete translated anchor-cluster preference accuracy on 2,800 examples (400 clusters × 7 languages) per model. ORPO is absent from the Gemma block (Section 3).

Method	Avg	Macro	Worst	Gap	EN	FR	IT	PT	HI	AR	ZH
Qwen3-4B											
Non-LARA baselines											
Base	53.5	53.5	51.2	-1.1	52.5	51.2	53.0	51.7	55.0	51.7	59.0
SFT	55.3	55.3	51.5	-0.3	55.0	54.0	51.5	53.8	57.5	53.2	62.0
DPO	55.1	55.1	52.2	**0.1**	55.2	52.8	52.2	52.8	57.5	55.0	60.5
ORPO	56.2	56.2	50.5	-2.6	54.0	55.5	50.5	55.8	60.2	56.5	61.3
LARA variants											
LARA w/o calib.	**65.4**	**65.4**	61.8	3.0	**68.0**	**67.0**	64.0	**65.8**	**66.8**	64.5	61.8
LARA	65.3	65.3	61.8	2.8	67.8	65.0	**65.2**	**65.8**	64.2	**67.5**	61.8
LARA+DRO	**65.4**	**65.4**	**63.7**	2.5	67.5	65.2	63.7	64.0	64.8	67.0	**65.2**
LARA+anchor	63.0	63.0	61.3	-0.8	62.3	61.3	64.0	63.7	66.0	62.3	61.3
LARA+DRO+anchor	62.5	62.5	61.0	-0.9	61.8	62.0	62.3	63.2	65.5	61.8	61.0
Aya Expanse 8B											
Non-LARA baselines											
Base	54.9	54.9	52.0	2.1	56.8	52.0	53.0	52.0	54.5	55.8	60.5
SFT	56.4	56.4	54.2	2.1	58.2	55.0	55.5	55.8	54.2	55.5	60.8
DPO	55.9	55.9	52.5	2.5	58.0	52.5	54.2	53.8	55.0	56.5	61.3
ORPO	57.5	57.5	55.8	1.4	58.8	56.5	55.8	56.8	56.5	57.2	61.3
LARA variants											
LARA w/o calib.	64.8	64.8	63.2	1.1	65.8	63.2	63.5	64.0	67.5	65.5	64.0
LARA	67.2	67.2	**65.5**	0.3	**67.5**	68.8	65.5	**68.8**	67.8	67.0	**65.5**
LARA+DRO	**67.6**	**67.6**	**65.5**	**-0.1**	**67.5**	**69.5**	66.0	68.2	**68.0**	**68.2**	**65.5**
LARA+anchor	65.2	65.2	62.7	0.7	65.8	62.7	**66.2**	64.5	66.5	65.8	64.8
LARA+DRO+anchor	65.1	65.1	62.5	0.8	65.8	62.5	66.0	64.8	66.8	65.2	64.5
Gemma-4-12B-it											
Non-LARA baselines											
Base	51.2	51.2	47.0	-4.6	47.2	47.0	50.2	49.8	55.0	54.5	54.5
SFT	55.0	55.0	51.2	-2.6	52.8	51.2	53.5	52.5	59.0	57.0	59.0
DPO	51.1	51.1	48.0	-3.6	48.0	48.2	48.8	49.5	55.0	53.2	54.8
LARA variants											
LARA w/o calib.	63.6	63.6	60.5	2.0	65.2	63.5	60.5	62.0	63.5	64.5	**65.8**
LARA	**66.5**	**66.5**	62.3	**0.0**	66.5	**68.8**	**66.5**	**67.5**	65.5	**68.2**	62.3
LARA+DRO	66.4	66.4	**63.7**	1.3	**67.5**	68.0	64.8	65.0	**67.8**	**68.2**	63.7
LARA+anchor	62.5	62.5	60.5	0.8	63.2	62.5	63.0	62.0	63.5	63.0	60.5
LARA+DRO+anchor	62.8	62.8	61.0	1.2	63.7	63.5	64.8	61.0	64.0	61.0	61.3

On Qwen, LARA w/o calib. and LARA+DRO lead at 65.4% average anchor accuracy, with LARA+DRO taking the best worst-language score of 63.7%. On Aya, LARA+DRO is strongest overall, reaching 67.6% average with a near-zero gap of -0.1, and on Gemma, where ORPO has no usable evaluation, LARA leads at 66.5% average with LARA+DRO best on worst-language (63.7%). On every backbone the unnormalized baselines remain in the low-to-mid 50s on anchor accuracy, with SFT and ORPO reaching at most 57.5% and no unnormalized Gemma baseline exceeding 55.0%, while LARA and LARA+DRO move anchor accuracy into the low-to-high 60s across all seven languages. The explicit anchor-loss variants are not the highest-accuracy anchor rows on any model, which is why the main text treats anchoring as a regularizer for balance rather than the primary accuracy driver.

On Qwen, LARA+DRO improves over DPO by $+10.3$ average and $+11.5$ worst-language points. On Aya the improvement is +11.7 average and +13.0 worst-language points, and on Gemma LARA gains +15.4 average and +14.3 worst-language points over DPO. Because each anchor cluster contains matched translations of the same semantic preference, these gains indicate that a normalized and per-language calibrated margin helps the model make the same preference decision across languages. The gaps here, however, are less reassuring than the held-out gaps. LARA and LARA+DRO reach 0.0 and 1.3 on Gemma and 0.3 and -0.1 on Aya, yet on Qwen every LARA variant except the

two anchor-bearing ones runs $+2.5$ to $+3.0$ in favor of English. Because all clusters are English-sourced, we read that Qwen column as a residual advantage for English. The remaining two diagnostics move from anchor *accuracy* to anchor *agreement*, testing whether calibration changes the structure of the preference signal rather than only its mean.

F.2 Held-Out Language Gap

Figure 17 plots the RQ3 language-balance result of Section 3.3 on the held-out test set. All LARA variants substantially reduce the language gap relative to DPO (-2.4) and ORPO (-2.0), with LARA reaching essentially zero (-0.0). The anchor-bearing variants, by contrast, lie slightly further from zero, which reflects the 2.6 points of average accuracy that LARA+anchor exchanges for cross-lingual agreement.

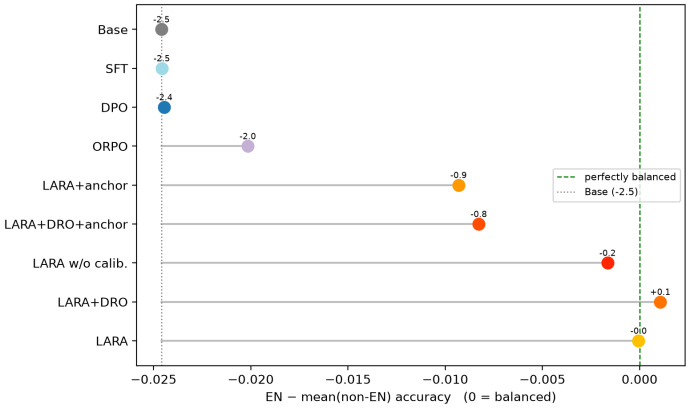


Figure 17: Language gap on Gemma-4-12B-it, namely English accuracy minus mean non-English accuracy. Values closer to zero indicate more balanced accuracy.

F.3 Cross-Lingual Transfer by Language Pair

Figure 18 measures, for each language pair, the share of translated anchor clusters on which the two languages agree on the margin sign. Under SFT, agreement ranges from 0.56 to 0.73 and decays for distant-script pairs such as French and Chinese at 0.56. Under LARA+DRO+anchor, by contrast, agreement rises to between 0.77 and 0.86 with gains of $+0.10$ to $+0.25$ on every pair, and the largest improvements appear exactly on the cross-script pairs. The anchor term thus makes the *same* preference decision travel across languages, which is the mechanism behind the anchor-cluster gains in Table 7.

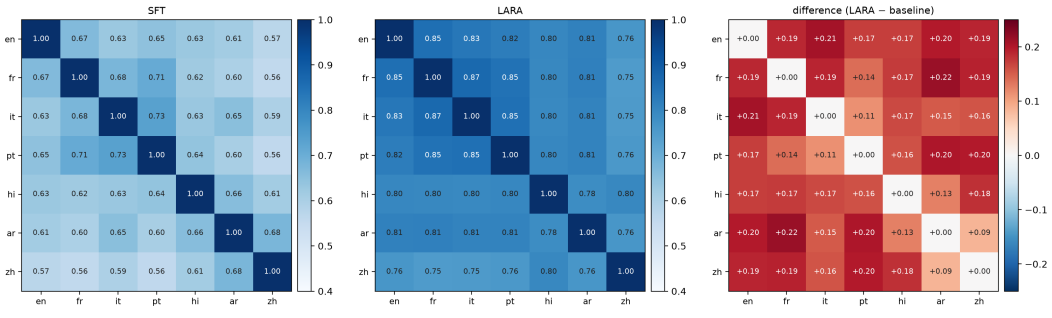


Figure 18: Cross-lingual sign agreement on anchor clusters (Gemma-4-12B-it) for SFT (left), LARA+DRO+anchor (middle), and their difference (right).

G GENERATION LENGTH STUDY

This section reports the free-form half of RQ4. This analysis is deliberately secondary to the preference evaluation, since it decodes 56 prompts (8 per language) per model rather than scoring 8,514 held-out pairs, and its purpose is to detect a different failure mode, degenerate or inflated generations, rather than to rank methods. Table 8 reports the response-length statistics for every method on all three backbones, which is the quantity that could otherwise explain the preference gains. One limitation bounds what this table demonstrates, however. Decoding is capped at 256 new tokens (Appendix I.3) and the mean token counts fall between 245 and 256, which means that most generations reach the cap and the length distribution stays censored from above. Within that ceiling, the table can detect a method that becomes markedly *shorter* or degenerate, and can rule out large verbosity differences below the cap, without measuring how much longer an unconstrained generation would have been. A study with a larger budget would test the verbosity question properly, and we therefore report this one as a screen rather than as decisive evidence.

Table 8: Generation length statistics (mean/median graphemes, mean tokens) for the 56-prompt study on all three models.

	Qwen3-4B			**Aya Expanse 8B**			**Gemma-4-12B-it**		
Method	Mean gr.	Med. gr.	Tokens	Mean gr.	Med. gr.	Tokens	Mean gr.	Med. gr.	Tokens
Reference									
Base	1109	1198	256	816	961	255	827	921	256
Non-LARA baselines									
SFT	1122	1213	256	804	899	253	831	934	255
DPO	1094	1189	256	816	984	254	823	916	256
ORPO	1107	1195	256	767	788	245	—	—	—
LARA variants									
LARA w/o calib.	1083	1174	256	780	769	251	814	865	255
LARA	1085	1200	256	755	843	250	824	889	256
LARA+DRO	1085	1193	256	751	832	251	822	893	255
LARA+anchor	1090	1196	256	771	891	253	817	872	255
LARA+DRO+anchor	1098	1212	256	766	866	253	826	908	256

The verbosity statistics highlight a key model difference. On Qwen, response lengths are uniform across methods (1,083 to 1,122 graphemes), which confirms that verbosity differences do not drive the Qwen preference gains. On Aya, the LARA variants run from 751 to 780 graphemes, below DPO (816) and straddling ORPO (767). On Gemma, finally, all analyzed methods stay close to the base length (814 to 831 graphemes), and the retrained ORPO checkpoint is absent because the generation study predates it (Section 3). No LARA variant exceeds the DPO response length by more than 0.4%, and on Aya the variants run 4 to 8% shorter. Within this decoding budget the LARA variants therefore match DPO in length to within half a percent, and longer answers do not explain the preference gains.

H QUALITATIVE AND ERROR ANALYSIS

This gallery expands Table 4 of Section 3.5 with further held-out pairs from the Gemma-4-12B-it evaluation. All decisions use the common length-normalized policy-score rule of Section 3. The reported margin is the grapheme-normalized score difference, where a positive value means that the human-preferred response ranks first, and LARA margins are additionally in per-language standard-deviation units. We quote French, Italian, and Portuguese pairs verbatim, while Arabic, Hindi, and Chinese pairs appear in English translation because the Latin font of the paper cannot render those scripts, and the original-script texts appear in the released evaluation files. We truncate excerpts at the marker "[. . .]".

We group the eleven examples by the failure mode they illustrate, in the order the mechanisms appear in the paper. Length traps where the human prefers the shorter response come first (Examples 1 to 3), followed by substance-over-brevity pairs where the human prefers the longer one (Examples 4 and 5), then rescues in the three non-Latin-script languages (Examples 6 to 9), and finally, for balance, cases where LARA is wrong and most baselines are right (Examples 10 and 11). Taken together, the first two groups are the per-example form of the length-bias diagnostic in Figure 3, since a method that decides by length must fail one group or the other.

H.1 Length Traps, Where the Human Prefers the Shorter Response

The following pairs probe whether a method follows verbosity or content.

Example 1. French (length trap)		
Prompt		Trouve des idées créatives pour un activisme écologique dans un roman pour jeunes adultes.
Preferred ch)	*(2,931*	"Voici quelques idées créatives pour un activisme écologique dans un roman pour jeunes adultes : 1. **Créer un personnage principal qui est un activiste écologique** : Le personnage principal pourrait être un jeune adulte qui est passionné par la protection de l'environnement et qui décide de prendre des mesures pour lutter contre la pollution, la déforestation ou le changement climatique. 2. **Inclure des thèmes écologiques dans l'intrigue** : L'intrigue du roman pourrait tourner autour d'une crise écologique, telle qu'une marée noire, une catastrophe nucléaire ou une sécheresse [. . .]" The answer builds novel material around characters and plot, as asked.
Rejected ch)	*(3,566*	"1. **Organiser un événement de nettoyage de plage ou de parc** : Les personnages du roman pourraient organiser un événement de nettoyage de plage ou de parc pour sensibiliser les gens à l'importance de protéger l'environnement. 2. **Créer un jardin communautaire** : Les personnages pourraient créer un jardin communautaire pour promouvoir l'agriculture durable [. . .]" Longer, yet a list of real-world actions rather than story ideas.
Decisions		Base × (-0.11) SFT × (-0.09) DPO × (-0.11) ORPO × (-0.01) LARA w/o calib. ✓ LARA ✓ $(+0.01)$ LARA+DRO+anchor ✓ $(+0.18)$

Example 2. Portuguese (length trap)		
Prompt		Qual é o melhor projetor portátil para noites de cinema ao ar livre?
Preferred ch)	*(1,222*	"Existem vários projetores portáteis excelentes para noites de cinema ao ar livre. Aqui estão algumas opções populares: 1. **Anker Nebula Capsule**: Um projetor portátil compacto e leve, com resolução Full HD e autonomia de bateria de até 4 horas. 2. **Epson EF-100**: Um projetor portátil com resolução Full HD e tecnologia de iluminação laser, que oferece uma imagem nítida e vibrante. 3. **Optoma ML750**: Um projetor portátil com resolução Full HD e tecnologia de iluminação LED [. . .]" Named products with specifications.
Rejected ch)	*(1,534*	"Para noites de cinema ao ar livre, é importante escolher um projetor portátil que seja leve, fácil de configurar e ofereça boa qualidade de imagem, mesmo em ambientes com alguma luz. Aqui estão algumas características e opções a considerar: 1. **Brilho (Lumens)**: Quanto mais lumens, melhor a visibilidade da imagem em ambientes com luz. Para uso ao ar livre, procure por projetores com pelo menos 2.500 lumens [. . .]" Longer, and it lists generic buying criteria before any recommendation.
Decisions		Base × (-0.14) SFT × (-0.03) DPO × (-0.13) ORPO ✓ $(+0.03)$ LARA w/o calib. ✓ LARA ✓ $(+0.60)$ LARA+DRO+anchor ✓ $(+0.34)$

Example 3. Italian (length trap)		
Prompt		Scrivi una sceneggiatura biografica su un personaggio storico e sul suo ruolo in una grande invenzione.
Preferred ch)	*(1,212*	"**Titolo: La vita di Guglielmo Marconi e la nascita della radio**. **Scena 1: L'infanzia di Marconi**. La sceneggiatura si apre con Guglielmo Marconi, un giovane ragazzo italiano, che cresce in una famiglia di agricoltori. Marconi è affascinato dalle scoperte scientifiche e trascorre ore a leggere e sperimentare. **Scena 2: La scoperta della radio**. Marconi scopre le ricerche di James Clerk Maxwell e Heinrich Hertz sulla propagazione delle onde elettromagnetiche. Egli intuisce il potenziale di queste onde per la comunicazione a distanza [. . .]" The answer supplies concrete scenes that advance the story.
Rejected ch)	*(1,456*	"**Titolo: La vita di Guglielmo Marconi e la nascita della radio**. **Scena 1: L'infanzia di Marconi**. La sceneggiatura si apre con Guglielmo Marconi, un giovane italiano con una passione per l'elettricità e la fisica. Marconi cresce in un ambiente che incoraggia la sua curiosità e il suo desiderio di innovazione [. . .]" The same outline restated at greater length with less scene detail.
Decisions		Base × (-0.01) SFT × (-0.02) DPO × (-0.02) ORPO × (-0.00) LARA w/o calib. ✓ LARA ✓ $(+0.03)$ LARA+DRO+anchor ✓ $(+0.11)$

H.2 Substance over Brevity, Where the Human Prefers the Longer Response

The same methods that pick verbose answers above also discard thorough ones here, which confirms that they track length rather than quality. LARA, by contrast, tracks the preferred content in both directions.

Example 4. Italian (substance)		
Prompt		Sto organizzando un matrimonio in una località lontana e vorrei stabilire budget e tempistiche realistici per l'evento.
Preferred ch)	*(1,560*	"Per stabilire un budget e tempistiche realistici per il tuo matrimonio in una località lontana, considera i seguenti passaggi: 1. **Definisci il numero degli ospiti**: Questo influenzerà direttamente il budget per il cibo, le bevande e la location. 2. **Scegli la data**: Le stagioni e i giorni della settimana possono influenzare i costi. 3. **Ricerca location**: Considera non solo il costo della location, ma anche le spese di viaggio e alloggio per te e i tuoi ospiti. 4. **Cibo e bevande** [. . .]" A step-by-step budgeting plan.
Rejected (616 ch)		"Un approccio olistico alla pianificazione del matrimonio può aiutare a creare un'esperienza unica e memorabile per gli sposi e gli ospiti. Considerare non solo il budget e le tempistiche, ma anche l'impatto ambientale e sociale dell'evento può portare a scelte più sostenibili e responsabili [. . .]" Short and abstract, without any budget steps.
Decisions		Base × (-0.12) SFT × (-0.05) DPO × (-0.15) ORPO × (-0.01) LARA w/o calib. ✓ LARA ✓ $(+5.34)$ LARA+DRO+anchor ✓ $(+2.79)$

Example 5. Chinese, translated (substance)	
Prompt	What are the best surf spots on Fernando de Noronha?
Preferred (223 ch)	"Brazil's Fernando de Noronha has several famous surf spots. The best known are 1. Cacimba do Padre, a reef break with powerful waves, 2. Bode, a point break with steady waves, 3. Meio, a reef break with strong waves, and 4. Conceição, a gentle beach break that is ideal for beginners. These spots are famous for their clear water and coral reefs and suit surfers of every level."
Rejected (184 ch)	"Fernando de Noronha is a UNESCO World Heritage site and a protected marine park with some of the world's most pristine surf. The best season is December to March, when the swell peaks. Famous spots include Cacimba do Padre, Bode, and Meio [...] note, however, that surfing there requires a permit and surfer numbers are capped to protect the environment." Background and caveats, naming spots only in passing.
Decisions	Base × (-0.27) SFT × (-0.08) DPO × (-0.25) ORPO ✓ ($+0.13$) LARA w/o calib. ✓ LARA ✓ ($+7.79$) LARA+DRO+anchor ✓ ($+4.33$)

H.3 Non-Latin-Script Rescues (Arabic, Hindi, Chinese)

Pairs from the three languages written in a non-Latin script, namely Arabic, Devanagari, and Han (Section 2.2). In each case every baseline ranks the pair incorrectly.

Example 6. Arabic, translated (rescue)	
Prompt	How can a physics degree lead to a materials-science career in the UK?
Preferred (1,178 ch)	"To leverage a physics degree into a materials-science career in the UK, consider: 1. **Postgraduate study**, a Master's or PhD in materials science for specialist knowledge. 2. **Relevant coursework**, covering solid-state physics, thermodynamics, and materials properties during your degree. 3. **Research experience**, meaning research projects or internships focused on materials science. 4. **Professional certification**, for example Chartered Engineer or Chartered Physicist through IOM3. 5. **Networking**, so join professional bodies, attend conferences, and connect with materials scientists in the UK [...]"
Rejected (817 ch)	"For physics graduates wishing to pursue a materials-science career in the UK, it is essential to appreciate the interdisciplinary nature of the field and be prepared to acquire new skills. Materials science combines physics, chemistry, and engineering to design and develop new materials, so physics graduates may need additional courses or workshops on materials chemistry, mechanical properties, and manufacturing processes. They can also explore emerging areas such as nanotechnology, biomaterials, or energy storage [...]" A reflection with no actionable steps.
Decisions	Base × (-0.07) SFT × (-0.04) DPO × (-0.06) ORPO × (-0.01) LARA w/o calib. ✓ LARA ✓ ($+4.16$) LARA+DRO+anchor ✓ ($+3.03$)

Example 7. Arabic, translated (rescue)	
Prompt	What are the best museums in Reims to learn about sparkling wine?
Preferred (545 ch)	"The best-known champagne museums in Reims include: 1. the **Veuve Clicquot museum**, which explores the history of the house; 2. Taittinger's **Le Fond de Bourse**, offering tours and tastings focused on the production process; 3. **Maison Mumm**, an interactive look at champagne production and the Mumm brand; 4. the **Museum of Fine Arts**, which, while not exclusively about champagne, exhibits the region's history and culture, including champagne production. Together they give a comprehensive picture of champagne's history, production, and cultural importance in Reims."
Rejected (540 ch)	"For a deeper understanding of the champagne industry and its impact on the local community, please visit the Museum of Art and History in Reims, which has a section devoted to the history of champagne production in the region. Its exhibits show champagne's social and economic importance, including its role in shaping the city's architecture, culture, and traditions [...]" Redirects to a single general museum.
Decisions	Base × (-0.20) SFT × (-0.19) DPO × (-0.19) ORPO × (-0.10) LARA w/o calib. ✓ LARA ✓ ($+4.10$) LARA+DRO+anchor ✓ ($+2.69$)

Example 8. Hindi, translated (rescue)	
Prompt	Suggest what to see and do on a weekend break in Montréal.
Preferred (439 ch)	"For a weekend break in Montréal you can visit: 1. **Old Montréal (Vieux-Montréal)**, a historic district famous for its beautiful architecture, street artists, and restaurants. 2. **The Montréal Tower**, with a view over the whole city. 3. **Jean-Talon Market**, known for local produce and food. 4. **The Montréal Botanical Garden**, with plants and flowers of many kinds. 5. **Place des Arts**, a cultural center famous for art, music, and dance performances. Beyond these, simply wandering Montréal's streets [...]"
Rejected (237 ch)	"During a weekend break in Montréal you can head to the city's outskirts and enjoy nature. In nearby Mount Royal Park you can trek and hike and enjoy splendid views of the city. You can also cycle along the Lachine Canal and take in the city's historical and cultural significance." Covers only one kind of activity.
Decisions	Base × (-0.09) SFT × (-0.01) DPO × (-0.08) ORPO × (-0.07) LARA w/o calib. ✓ LARA ✓ ($+7.79$) LARA+DRO+anchor ✓ ($+2.96$)

Example 9. Chinese, translated (rescue)	
Prompt	I am a writer commissioned to write an article about a new sustainable technology in my community, and I would like help writing about its benefits and challenges.
Preferred (74 ch)	"I'd be glad to help. Which sustainable technology are you implementing? Is it related to energy, water, waste management, or transport? Knowing this will help me give you more specific material, including its advantages and challenges." A clarifying question before drafting.
Rejected (226 ch)	"Sustainable technology is not only about reducing our environmental impact; it also involves advancing social justice and equity. In many communities, low-income and marginalized residents are disproportionately affected by pollution and lack access to clean energy and green space. Implementing sustainable technology can help address this by providing affordable clean-energy options, improving air and water quality, and fostering community participation [...]" A generic essay that never asks which technology is meant.
Decisions	Base × (-0.40) SFT × (-0.57) DPO × (-0.44) ORPO × (-0.07) LARA w/o calib. ✓ LARA ✓ ($+0.29$) LARA+DRO+anchor ✓ ($+0.72$)

H.4 Error Analysis

For balance, two pairs where LARA is wrong and most baselines are right. Both involve nearly interchangeable English answers whose preference hinges on fine-grained detail rather than on content or length.

Example 10. English (LARA failure)	
Prompt	What's the best way to explore the historic center of Ouro Preto?
Preferred (460 ch)	"The best way to explore the historic center of Ouro Preto is on foot, allowing for a leisurely pace to absorb the colonial architecture, visit numerous churches, and take in the picturesque streets and alleys. This method provides an intimate connection with the town's history and culture, enabling visitors to discover hidden gems and unexpected viewpoints. The town's compact size makes walking a practical and enjoyable way to experience its rich heritage."
Rejected (269 ch)	"The best way to explore the historic center of Ouro Preto is on foot. Wear comfortable shoes and wander the cobblestone streets to discover colonial architecture, historic churches, and picturesque alleys. Guided tours are also available for a more in-depth experience." A tighter version of the same advice, and LARA's per-unit score favors this denser variant.
Decisions	Base ✓ $(+0.06)$ SFT ✓ $(+0.01)$ DPO ✓ $(+0.06)$ ORPO ✓ $(+0.02)$ LARA w/o calib. × LARA × (-0.22) LARA+DRO+anchor × (-0.33)

Example 11. English (LARA failure)	
Prompt	I'm part of a college football team and I need help drafting an encouragement note to send to our coach before our big rivalry rematch against University of Michigan.
Preferred (518 ch)	"This rematch is an opportunity to prove yourselves, to show that you've learned from past experiences and that you're stronger, faster, and more united than ever. Tell your coach that you believe in their strategy and in the team's ability to execute it flawlessly. Emphasize the importance of staying focused, of trusting in each other's skills, and of playing with heart and passion [...]" Coaches the student on what to say.
Rejected (241 ch)	"Dear Coach, We're ready to give it our all in the rematch against Michigan. Your leadership and guidance have prepared us for this moment. We're confident in our abilities and in each other. Let's bring home a win! Sincerely, [Your Team]" A ready-to-send note, and LARA prefers this compact finished draft.
Decisions	Base ✓ $(+0.03)$ SFT × (-0.05) DPO ✓ $(+0.02)$ ORPO ✓ $(+0.04)$ LARA w/o calib. × LARA × (-1.00) LARA+DRO+anchor × (-0.70)

Across the gallery, then, the pattern of Table 4 recurs. The baselines decide by likelihood and length, LARA decides by calibrated per-unit preference evidence, and the LARA errors concentrate on near-duplicate pairs in which the human signal itself remains subtle.

I Implementation Details and Hyperparameters

I.1 Models and Adaptation

The three backbones are the implementations from Huggingface with the following model name: `Qwen/Qwen3-4B`, `CohereLabs/aya-expanse-8b`, and `google/gemma-4-12B-it`. All methods train LoRA adapters on a frozen base model in bf16 with gradient checkpointing, while the Gemma base weights are additionally loaded in 4-bit (Dettmers et al., 2023) for the LARA, DPO, and SFT runs (the retrained ORPO run uses bf16 weights). Beyond adaptation, Table 9 lists all training hyperparameters, and every method uses the same prompt formatting, 2,048-token sequence limit, data split, and one-epoch budget on its backbone. Two settings are not shared across the two families and should be read as confounds when comparing a LARA variant with a baseline, namely the learning rate (5×10^{-6} for LARA variants against 1×10^{-5} for baselines) and the seed (13 against 42). Each setting follows the configuration under which that family was tuned, and we did not re-tune one family under the other's setting. Weight precision separates one run from the rest instead, since on Gemma the retrained ORPO baseline runs in bf16 where LARA, DPO, and SFT load the base weights in 4-bit.

I.2 Data and Splits

Preference pairs come from Community Alignment, with EN, FR, IT, PT, and HI native and AR and ZH added as machine-translated extensions, converted to pairwise form as described in Section 2 and split by conversation identifier into training (31,000 pairs, namely 5,000 each for EN, FR, IT, PT, and HI and 3,000 each for AR and ZH), a calibration split of at most 500 pairs per language (3,426 in total, namely 500 each for EN, PT, AR, and ZH and 474, 489, and 463 for FR, IT, and HI, which are the smaller native subsets), and the 8,514-pair held-out test set, whose per-language sizes are also unequal (1,500 each for EN, AR, and ZH and 902 to 1,143 for FR, IT, PT, and HI), which is why we report macro accuracy alongside the pooled average. The 400 translated anchor clusters (2,800 pairs) are built by translating English *training* pairs into the other six languages and

Table 9: Training hyperparameters. Shared values apply to every method on the backbone, and LARA-specific terms are active only in the variants that use them.

	Qwen3-4B	Aya Expanse 8B	Gemma-4-12B-it
Shared			
LoRA rank r / α / dropout		16 / 32 / 0.05	
Precision		bf16, gradient checkpointing	
Max sequence length		2,048 tokens	
Epochs / warmup ratio / weight decay		1 / 0.03 / 0.0	
Max gradient norm		1.0	
DPO/LARA temperature β		0.1	
Micro-batch $\times$ grad. accumulation	2×16	2×16	1×32
Learning rate (LARA variants)		5×10^{-6}	
Learning rate (SFT/DPO/ORPO)		1×10^{-5}	
Seed (LARA / baselines)		13 / 42	
4-bit base-model loading	no	no	LARA, DPO, SFT
LARA-specific			
Calibration floor $s_{\min}$		0.03	
Calibration pairs per language		up to 500 (3,426 total)	
Effective length $e(y)$		Unicode graphemes	
Anchor weight λ		0.03	
DRO temperature η / EMA ρ		1.0 / 0.9	

serve only the anchor term and the consistency diagnostics rather than a held-out set. Finally, we precompute reference log-probabilities for all splits once with the frozen base model and reuse the same values for all nine methods.

I.3 Evaluation Protocol

Held-out and anchor-cluster scoring uses the length-normalized policy-score rule of Section 3 (chunked scoring, batch size 4 and a 2,048-token limit). Grouped bootstrap intervals use 5,000 conversation-level resamples with seed 13, and the diagnostic figures use 800 resamples as stated in their captions. Where a median appears in both places, the two differ in the first decimal for that reason. The generation study decodes 56 prompts (8 per language) per model under identical decoding settings, greedily and with a budget of 256 new tokens, and measures response length in Unicode graphemes and model tokens. The cap binds for most prompts and bounds what the study can conclude (Appendix G).

I.4 Pipeline

A single driver script runs the following stages per backbone, namely data preparation and translation, reference scoring, calibration, training of all methods, held-out and anchor evaluation, bootstrap intervals, diagnostics, and the generation study. Downstream, separate plotting scripts regenerate every figure from the released per-example evaluation files. We generate all tables in this paper programmatically from the same files.

J Limitations

We cover three open multilingual models and seven languages, of which Arabic and Chinese are machine-translated extensions, and the generation study is a length comparison over 56 prompts rather than a human evaluation of quality. Three settings bound what the comparisons establish. The group-DRO weights are normalized within each micro-batch, which leaves the reweighting nearly or entirely inert at our sizes, and the Gemma LARA+DRO results therefore describe a second LARA run rather than a distinct objective (Section 2.5). The anchor clusters are all seeded from English training pairs, and the anchor term is therefore tested only in its English-to-other-languages form. The LARA and baseline families differ in learning rate and seed, and on Gemma the retrained ORPO run also differs in weight precision (Table 9), which family-level comparisons carry as a confound. Within those limits the results support one objective-level principle, namely that multilingual preference losses should place their margins on a comparable scale before aggregation.

K STATEMENTS

K.1 REPRODUCIBILITY STATEMENT

All reported numbers derive from released per-example evaluation files and deterministic scoring scripts. Section 3 defines the scoring rule exactly, and Appendix I lists every training and evaluation hyperparameter, model identifier, data split, and random seed, and describes the pipeline scripts that reproduce training, scoring, calibration, bootstrap intervals, and every figure and table from raw data. We trained each method once per backbone with a fixed seed, and statistical claims therefore rest on grouped-bootstrap resampling and rank-based tests over the 8,514 held-out pairs rather than on repeated training runs. Since a single run per method cannot separate an objective from run-to-run variance, the two places where a pair of our configurations is nominally the same objective are worth noting. On Gemma the reweighting is inert, which makes LARA+DRO and LARA the same objective, as it does LARA+DRO+anchor and LARA+anchor (Section 2.5). The runs in each pair share the seed of Table 9 yet still differ by 0.1 to 1.4 points, which is nondeterminism in training and scoring rather than an effect of the objective, and which therefore bounds the run-to-run noise on this setup. We will release code, configuration files, calibration statistics, and evaluation outputs, including those for the runs excluded from earlier drafts.

K.2 ETHICS STATEMENT

This work uses only publicly released resources, namely the Community Alignment preference dataset and open-weight pretrained models. We collected no new human-subject data and process no personally identifying information, and all prompts and responses quoted in the paper and appendix come from the public evaluation split. Two of the seven languages (Arabic and Chinese) are machine-translated extensions rather than native preference data, which may import translation artifacts and under-represent culturally specific preferences. We label them as such throughout and treat conclusions for these languages with corresponding caution. The method itself aims to reduce a fairness problem, namely the systematic under-alignment of lower-resourced languages under pooled preference objectives. Even then, calibrated preference optimization inherits whatever biases exist in the underlying preference annotations, and deployment does not remove the need for a separate safety evaluation in every language.

K.3 AI USE STATEMENT

We used generative AI tools in this work only as general-purpose assistants under our own supervision. We limited that use to polishing text we had already drafted, covering grammar, phrasing, and formatting, and to routine coding tasks such as plotting scripts and refactoring of experiment harnesses. At no point did we use them for research ideation, for deriving or checking the theoretical results, for generating experimental results or data, or for retrieving and summarizing related work. We reviewed all AI-assisted text and code, and every number we report comes from the released code that we ran ourselves.

L LICENSE

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, available at `https://creativecommons.org/licenses/by/4.0/`. The license permits copying, redistribution, and adaptation in any medium or format, including for commercial purposes, provided that appropriate credit is given, a link to the license is provided, and any changes are indicated. The same license applies to the code, configuration files, calibration statistics, and evaluation outputs released with this work. The license does not, however, extend to third-party material, and the backbone models and the preference corpus of Section 3 therefore remain governed by their own licenses and terms of use.