MORPHOLOGY-CONSISTENT PREFERENCE OPTIMIZATION FOR INDIC LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference optimization has become the standard practice for aligning large language models, yet every dominant objective scores whole responses or raw subword tokens. The mismatch is acute in morphologically rich Indic languages, where one grammatical unit can decide rejection while the rest of the response stays correct. Case suffixes, agreement and tense-aspect-mood markers, honorific forms, and script-sensitive vowel signs all carry that weight. Subword tokens do not fill that role either as one token may conflate a stem with a suffix, and one morpheme may split across tokens. A token-level objective therefore spreads its signal past the unit that decided the preference. We introduce **MorphDPO**, which moves the unit of alignment from a flat token sequence to a latent morpheme lattice. The resulting reward replaces the sequence-level Direct Preference Optimization (DPO) log-ratio with a normalized morpheme-lattice reward. The new reward is tokenizer-agnostic, marginalizes over uncertain analyses, and reduces exactly to DPO under uniform weights. The underlying tilted-policy log-ratio keeps morphology weighting a valid change of policy geometry rather than an unnormalized token reweighting. Furthermore, we build three components, namely counterfactual partial-order preferences from rule-generated minimal pairs, a paradigm-neighborhood objective enforcing consistency across valid inflections, and a group-distributionally-robust objective over language-script-morphology groups. Across eight Indic languages on the Indic-specialized Sarvam-1 2B, MorphDPO attains the best composite accuracy (0.941) and win-rate (0.985) among preference objectives, with the highest morphology minimal-pair accuracy and a group gap tied with DPO for the smallest. The same lead holds on Llama-3.2 3B and Qwen3 1.7B, proving that morphology rather than the token is the right unit of preference alignment for morphologically rich languages.

1 INTRODUCTION

Direct Preference Optimization (DPO) reduced reinforcement learning from human feedback (Christiano et al., 2017; Ouyang et al., 2022) to one classification loss over chosen and rejected responses (Rafailov et al., 2023). Successor objectives vary the reference model, the labels, the noise model, and the granularity of the training signal (Meng et al., 2024; Hong et al., 2024; Ethayarajh et al., 2024; Wu et al., 2025; Zeng et al., 2024; Liu et al., 2025; Zhou et al., 2025), yet the unit stays the whole response or its subword tokens. Neither suits *morphologically rich* Indic languages, where acceptability turns on case, agreement, tense-aspect-mood (TAM) suffixation, honorificity, and sandhi. One wrong morpheme can decide the preference, a signal that sequence-level objectives dilute across hundreds of tokens and that reweighted "important tokens" cannot recover, since subword tokenizers optimize compression rather than grammar (Brahma et al., 2025).

Average win-rate averages away the rare morphology that decides acceptability, low-frequency inflections and uncommon honorific registers. Prior Indic work advances adjacent fronts through morphology-aware tokenization (Brahma et al., 2025), multilingual preference transfer (She et al., 2024), and retrieval for reward models (Ghosal et al., 2025), leaving the preference objective untouched. Our hypothesis is that alignment fails in the tail because morphology is not the optimization unit, and that objectives should be invariant to valid morphology and sensitive to invalid morphology.

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

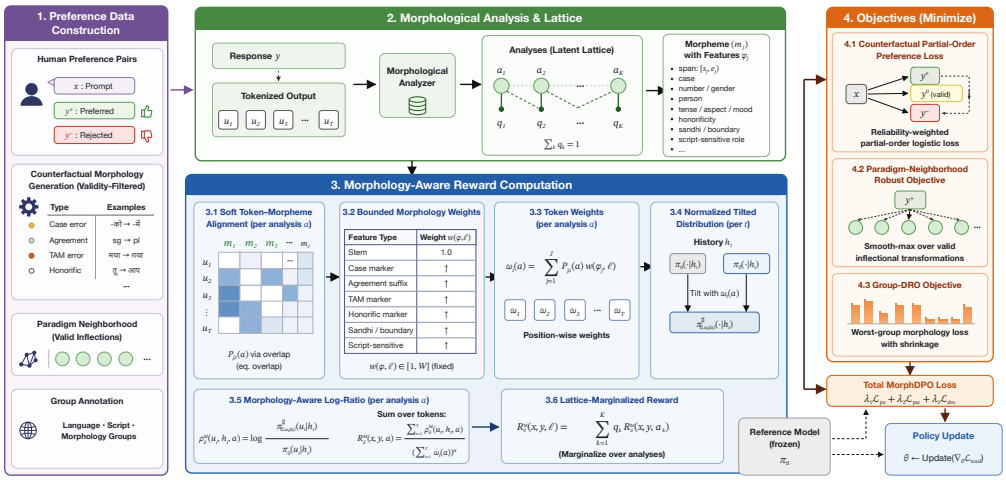


Figure 1: **Overview of MorphDPO.** (1) Preference pairs are augmented with rule-generated counterfactual corruptions, paradigm neighborhoods, and language-script-morphology group labels. (2) An analyzer maps each response to a lattice of analyses with posterior $q_\ell(a \mid y)$, each morpheme carrying a feature bundle ϕ_j. (3) Character-span overlap yields a soft token-morpheme alignment, bounded weights $w(\phi, \ell) \in [1, W]$ set position-wise $\omega_t(a)$, and the normalized tilted distribution $\pi^{\sharp}_{\theta,\omega_t(a)}$ defines a valid local log-ratio summed over tokens and marginalized over analyses into $R^M_\theta(x, y, \ell)$. (4) The reward enters a reliability-weighted partial-order loss, a paradigm smooth-max, and a group-DRO objective, minimized against a frozen reference. Solid arrows carry data and information flow, dashed arrows from the frozen reference carry no gradient, and thin dashed arrows mark supervision and partial-order relations.

Addressing these drawbacks, in **MorphDPO** (Figure 1) we score a response over a latent morpheme lattice from a morphological analyzer rather than a flat token sequence, with each morpheme softly aligned to model tokens by character-span overlap. Every token's contribution passes through a normalized tilted policy, which keeps the reward a genuine log-ratio of normalized distributions and a strict generalization of DPO. Three components build on that reward, namely counterfactual partial orders from rule-generated corruptions, a paradigm-neighborhood robust objective, and a group distributionally robust objective over language-script-morphology groups (Sagawa et al., 2020). On an Indic-specialized 2B backbone, MorphDPO reaches the best composite accuracy (0.9413) and overall win-rate (0.9848) among preference objectives, the highest morphology minimal-pair accuracy (0.9805 against 0.9732 for DPO), and a group gap of 0.2340 at a worst-group win-rate of 0.7660, each of the latter two tied with DPO. Per corruption type MorphDPO leads on wrong case (0.9865), gender (0.9779), and number agreement (0.9372), the only method in our comparison that improves ranking without losing morphological correctness. The lead is also consistent in Llama-3.2 (3B) and Qwen3 (1.7B), where MorphDPO outperforms the strongest shared baseline by 3.3 and 2.8 points of composite accuracy and 4.3 points of worst-group win-rate. The contributions are as follows.

- **Normalized Morpheme-Lattice Preference Optimization:** We derive a morphology-aware preference reward from a normalized tilted-policy log-ratio over a latent morpheme lattice, tokenizer-agnostic, marginalized over uncertain analyses, and exactly DPO under uniform weights, which avoids a well-known mathematical pitfall, the unnormalized reweighting of tokens.
- **Counterfactual, Paradigm, and Worst-Group Supervision:** We turn rule-generated morphological corruptions into causal minimal-pair supervision through a reliability-weighted partial-order loss, and we add a smooth-max objective over valid paradigm transformations and a shrinkage-stabilized group-DRO objective, each an upper bound on the worst-case risk it targets.
- **Empirical Validation on an Indic Backbone:** We evaluate eight preference objectives on eight Indic languages under one protocol and repeat six of them on three backbones, isolate the "morphology tax" that every baseline incurs, and attribute, on the Indic-specialized backbone, the tail robustness of MorphDPO to reward normalization through controlled ablations that also bound their own run-to-run spread.

2 METHOD

MorphDPO defines the implicit reward over latent morpheme analyses and couples that reward with counterfactual morphology preferences, paradigm-consistency regularization, and worst-group optimization. The objective is *tokenizer-agnostic* and *probabilistically normalized*, and the objective *reduces exactly to DPO* when morphology is uninformative. The following subsections describe each component in detail.

2.1 PRELIMINARIES AND MORPHEME LATTICES

Let $\mathcal{D}_h = \{(x_i, y_i^+, y_i^-, \ell_i)\}_{i=1}^N$ be a preference dataset of prompts, preferred and rejected responses, and languages, with π_θ the trainable policy and π_0 a frozen reference policy, the pretrained backbone itself in our experiments. For $y = (u_1, \dots, u_T)$ with history $h_t = (x, u_{<t})$, DPO scores y through $r_\theta(x, y) = \sum_{t=1}^T \log \frac{\pi_\theta(u_t|h_t)}{\pi_0(u_t|h_t)}$ and minimizes

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\big(\beta \left[r_\theta(x, y^+) - r_\theta(x, y^-)\right]\big), \tag{1}$$

where σ is the logistic function and $\beta > 0$ controls the deviation from π_0. MorphDPO replaces this flat score with a morphology-aware reward $R_\theta^M(x, y, \ell)$. For a response y in language ℓ, a morphological analyzer returns candidate analyses $\mathcal{A}_\ell(y) = \{a_1, \dots, a_K\}$ with posterior $q_\ell(a \mid y)$, which collapses to a point mass when one analysis is available. Marginalizing over this posterior, rather than committing to a single hard segmentation, matters because Indic morphology is frequently ambiguous. Each analysis decomposes y into morphemes $m_j = (s_j, e_j, \phi_j)$ carrying a character span and a feature bundle ϕ_j from a universal schema (Batsuren et al., 2022) that covers case, agreement, TAM, honorificity, suffix-stack position, sandhi status, and orthographic role.

2.2 BOUNDED MORPHOLOGY WEIGHTS

Each feature bundle ϕ receives a bounded weight $w(\phi, \ell) \in [1, W]$, whose lower bound keeps every token in play and whose upper bound prevents instability on rare features. A linguistic prior assigns weight one to stems and larger weights to case markers, agreement and TAM suffixes, honorific markers, sandhi boundaries, and vowel signs, and a frequency-aware variant adds severity and inverse-frequency terms (Appendix A). We hold w *fixed* during optimization, because freely learned weights would adapt to make the objective easier without improving the model. To reach model tokens without a new tokenizer, let u_t occupy span $[p_t, p_{t+1})$ and $O_{j,t}(a) = |[s_j, e_j) \cap [p_t, p_{t+1})|$ its overlap with m_j under a. The soft assignment and position-wise weight are

$$P_{j,t}(a) = \frac{O_{j,t}(a)}{\sum_{j'=1}^J O_{j',t}(a) + \epsilon}, \qquad \omega_t(a) = \sum_{j=1}^J P_{j,t}(a)\, w(\phi_j, \ell), \tag{2}$$

with $\epsilon > 0$ for stability. The many-to-many assignment mixes weights for a token spanning several morphemes and spreads a split morpheme over those positions.

2.3 NORMALIZED MORPHOLOGY-AWARE REWARD

A naive morphology-aware reward would form the weighted sum $\sum_t \omega_t(a) \log \frac{\pi_\theta(u_t|h_t)}{\pi_0(u_t|h_t)}$, which is undesirable, because an arbitrarily weighted sum of token log-ratios is not in general the log-ratio of normalized distributions, and the DPO derivation breaks down. MorphDPO instead tilts the reference toward the policy and renormalizes over the vocabulary $\mathcal{V}$, following the exponentially reweighted form that the Kullback-Leibler-regularized optimum takes (Peters & Schaal, 2007; Rafailov et al., 2023). With $\omega = \omega_t(a)$,

$$\pi^\sharp_{\theta,\omega}(v \mid h_t) = \frac{\pi_0(v \mid h_t)\, \exp\Big(\omega \log \frac{\pi_\theta(v|h_t)}{\pi_0(v|h_t)}\Big)}{Z_\theta(h_t, \omega)}, \quad Z_\theta(h_t, \omega) = \sum_{v \in \mathcal{V}} \pi_\theta(v \mid h_t)^\omega\, \pi_0(v \mid h_t)^{1-\omega}. \tag{3}$$

Because policy and reference are softmax distributions, the tilt is normalized for every bounded ω, and the token contribution becomes the *valid* local log-ratio

$$\rho_\theta^M(u_t, h_t, a) = \log \frac{\pi^\sharp_{\theta,\omega_t(a)}(u_t \mid h_t)}{\pi_0(u_t \mid h_t)} = \omega_t(a) \log \frac{\pi_\theta(u_t \mid h_t)}{\pi_0(u_t \mid h_t)} - \log Z_\theta(h_t, \omega_t(a)). \tag{4}$$

Subtracting $\log Z_\theta$ turns morphology weighting into a change of local policy geometry rather than an unconstrained rescaling of losses, and the normalization is local in a precise sense. For any fixed weight sequence $\omega_{1:T}$, the product $\prod_t \pi^\sharp_{\theta,\omega_t}(\cdot \mid h_t)$ is a valid autoregressive distribution over responses whose log-ratio to $\pi_0(y \mid x)$ is exactly $\sum_t \rho^M_\theta$, which makes the tilt a change of geometry conditional on the analysis rather than a single global policy. Controlling for length and morphology density, the analysis-specific reward and its marginalization are

$$R^M_\theta(x, y, a) = \frac{\sum_{t=1}^T \rho^M_\theta(u_t, h_t, a)}{\left(\sum_{t=1}^T \omega_t(a)\right)^\alpha}, \qquad R^M_\theta(x, y, \ell) = \mathbb{E}_{a\sim q_\ell(a|y)}\left[R^M_\theta(x, y, a)\right], \tag{5}$$

with $\alpha \in [0, 1]$, where $\alpha = 0$ recovers an unnormalized DPO-like score and $\alpha = 1$ an average morphology-weighted score. Every full-scale run uses $\alpha = 1$, and $\alpha = 0$ is the "unnormalized" variant ablated in Table 2. Under a point-mass posterior with $\omega_t(a) = 1$ and $\alpha = 0$ we have $Z_\theta(h_t, 1) = 1$ and $R^M_\theta = \log \frac{\pi_\theta(y|x)}{\pi_0(y|x)}$, reducing the method exactly to DPO (Proposition 1).

2.4 Reliability-Weighted Preference Loss

For an edge $e = (x, y^+, y^-, \ell, \gamma_e, \delta_e, b_e)$ with label reliability $\gamma_e \in [0.5, 1]$, margin $\delta_e \geq 0$, and morphology group b_e, we define the margin $m_\theta(e) = R^M_\theta(x, y^+, \ell) - R^M_\theta(x, y^-, \ell)$ and the reliability-weighted pairwise loss

$$\ell(e; \theta) = -\gamma_e \log \sigma(\beta[m_\theta(e) - \delta_e]) - (1 - \gamma_e) \log \sigma(-\beta[m_\theta(e) - \delta_e]). \tag{6}$$

Fully trusted reference pairs take $\gamma_e = 1$ and $\delta_e = 0$, recovering the usual logistic loss equation 1, while $\gamma_e < 1$ treats an automatically generated edge as partially reliable and δ_e encodes the severity of the induced error. The selected configuration sets $\delta_e = 0$ on every edge (§3.1), which leaves γ_e as the only per-edge quantity active in our runs.

2.5 Counterfactual Morphology Preferences

To cover rare morphology, we build a counterfactual preference set from a library $\mathcal{T}_\ell$ of language-indexed transformations, each targeting one corruption type c such as wrong case, wrong agreement, incorrect TAM suffix, honorificity mismatch, word-order transposition, sandhi corruption, or vowel-sign error. Given a high-quality response y, the transformation produces $\tilde{y} = T_c(y)$, retained as $y \succ \tilde{y}$ only when a strict validity predicate holds,

$$V_c(x, y, \tilde{y}) = 1 \iff \mathrm{SemPres}(x, y, \tilde{y}) \wedge \mathrm{TargetErr}_c(\tilde{y}) \wedge \bigwedge_{c' \neq c} \mathrm{NoExtraErr}_{c'}(\tilde{y}). \tag{7}$$

The predicate requires semantic preservation, introduction of the target error, and no unintended additional errors, yielding $\mathcal{D}_c = \{(x, y, T_c(y), \ell, c, \gamma_c, \delta_c) : V_c = 1\}$. Equation equation 7 states the contract we ask of a transformation, and our generator approximates rather than certifies the contract, accepting a pair whenever the rule for c fires and the result passes a length sanity check (§3.1). The reliability γ_c is therefore a fixed per-type constant rather than a per-pair confidence, and δ_c is available for severity while every reported run leaves it at zero. Minimal pairs are an established diagnostic for grammatical knowledge (Warstadt et al., 2020), and because the two responses differ only in the targeted property, this construction turns them into causal training supervision (Theorem 2).

With several acceptable and corrupted variants for one prompt, we avoid a spurious total order over $\mathcal{E}_{\mathrm{PO}}(x) = \{(y^+_p, y^-_q) : y^+_p \in Y^+(x),\ y^-_q \in Y^-(x)\}$ by minimizing the mean pairwise loss $\mathcal{L}_{\mathrm{PO}}(x)$ over that edge set, leaving equally acceptable variants unordered while lowering corrupted outputs.

2.6 Paradigm-Neighborhood and Worst-Group Robustness

MorphDPO further enforces consistency across valid paradigm transformations. For an edge e, let $\mathcal{G}_\ell(e)$ collect the transformations that map the prompt and both responses into a new valid inflection while preserving the preference label, such as singular-to-plural, gender-frame, tense-frame, or case-frame changes, with $g(e)$ the transformed edge. The robust loss is a smooth maximum over that

neighborhood,

$$\mathcal{L}_{\mathrm{par}}(e) = \tau_g \log \sum_{g \in \mathcal{G}_\ell(e)} \exp\left(\frac{\ell(g(e);\theta)}{\tau_g}\right), \tag{8}$$

with temperature $\tau_g > 0$ and $\mathcal{L}_{\mathrm{par}}(e) \to \max_g \ell(g(e);\theta)$ as $\tau_g \to 0$. Optimizing against the hardest valid variant discourages appearing aligned on frequent inflections while degrading on rarer ones (Theorem 3). Average accuracy can mask systematic failures on low-resource languages, rare scripts, or infrequent morphology types, and we therefore optimize the tail with group distributionally robust optimization (group-DRO) (Sagawa et al., 2020). Each edge joins a group $b_e \in \mathcal{B}$ defined by language, script, feature type, corruption type, and resource level, with $\mathcal{E}_b = \{e : b_e = b\}$, $n_b = |\mathcal{E}_b|$, and $\bar{\mathcal{L}}_b(\theta) = n_b^{-1} \sum_{e \in \mathcal{E}_b} \ell(e;\theta)$. Because small groups yield high-variance estimates, we shrink each group loss toward its parent $\mathrm{pa}(b)$, obtained by dropping the most specific attribute, as $\widetilde{\mathcal{L}}_b = \frac{n_b}{n_b+\rho}\bar{\mathcal{L}}_b + \frac{\rho}{n_b+\rho}\bar{\mathcal{L}}_{\mathrm{pa}(b)}$ with strength $\rho > 0$, falling back to the global mean without a parent. We then minimize the smooth worst-group objective

$$\mathcal{L}_{\mathrm{DRO}}(\theta) = \eta \log \sum_{b \in \mathcal{B}} \exp\left(\frac{\widetilde{\mathcal{L}}_b(\theta)}{\eta}\right), \qquad \eta > 0, \tag{9}$$

which confines the worst-group loss, $\max_b \widetilde{\mathcal{L}}_b \le \mathcal{L}_{\mathrm{DRO}} \le \max_b \widetilde{\mathcal{L}}_b + \eta \log |\mathcal{B}|$ (Theorem 3), aligning training with the goal of improving the weakest groups.

2.7 Full Objective and Implementation

With $\mathcal{E}_{\mathrm{par}}$ the subset of edges with valid paradigm neighborhoods, the full objective combines morphology-aware pairwise optimization, counterfactual partial orders, paradigm robustness, and group-DRO:

$$\min_\theta \; \mathcal{J}_{\mathrm{MorphDPO}}(\theta) = \mathcal{L}_{\mathrm{DRO}}(\theta) + \lambda_{\mathrm{PO}}\, \mathbb{E}_x[\mathcal{L}_{\mathrm{PO}}(x)] + \lambda_{\mathrm{SFT}}\, \mathcal{L}_{\mathrm{SFT}}(\theta) + \frac{\lambda_{\mathrm{par}}}{|\mathcal{E}_{\mathrm{par}}|} \sum_{e \in \mathcal{E}_{\mathrm{par}}} \mathcal{L}_{\mathrm{par}}(e), \tag{10}$$

where the optional regularizer $\mathcal{L}_{\mathrm{SFT}}$ is a token-averaged negative log-likelihood with a small coefficient (Appendix A). Coefficients and the scale of the morphology weights are selected on a held-out validation split, the bound W stays fixed, and π_0 stays frozen.

Analyses, counterfactual edges, and paradigm neighborhoods are precomputed offline, and the only extra computation is $\log Z_\theta$, one logsumexp per target position that keeps the per-step cost within a small constant factor of DPO and that the selected full-scale configuration disables (§3.1). Without a usable segmentation, the method reduces locally to DPO (Appendix A). Appendix B proves exact DPO reduction under uniform weights, an $O(\varepsilon)$ perturbation under analyzer error, correct counterfactual labels, and smooth bounds on the paradigm and worst-group risks (Proposition 1, Theorems 1 to 3). In summary, MorphDPO moves the unit of preference optimization from a token sequence to a morpheme lattice, preserving the normalized structure of DPO while adding counterfactual partial orders, paradigm robustness, and worst-group optimization.

3 Experiments

Our evaluation is organized around three research questions as follows:

RQ1 *Ranking quality.* Does aligning preferences over morphological units improve overall preference-ranking quality, relative to sequence-level and token-level objectives?

RQ2 *Quality without a morphology tax.* Can a preference objective improve ranking quality *without* sacrificing morphological correctness, across individual corruption types and the weakest language-script-morphology groups?

RQ3 *Where do the gains come from?* Which design choices, among marginalizing over analyzer uncertainty, normalizing the morphology-weighted reward, and the exact tilted-policy normalization, are responsible for the observed behavior?

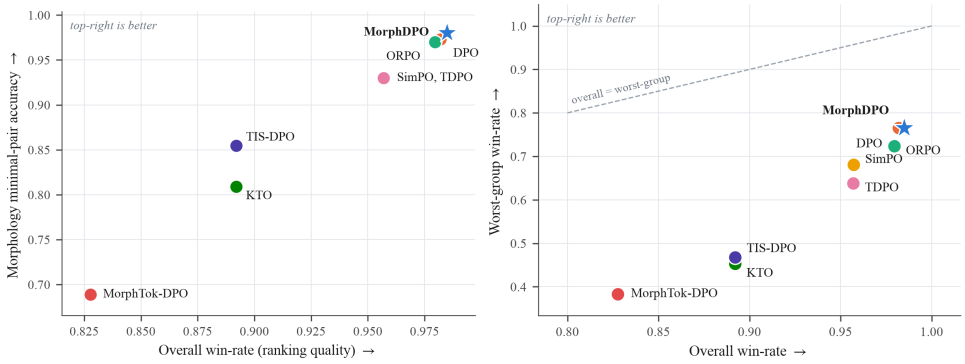


Figure 2: **No morphology tax. Left:** overall win-rate against minimal-pair accuracy. **Right:** overall against worst-group win-rate, with the dashed diagonal marking equality. MorphDPO (star) is the unique top-right point in both planes.

3.1 Experimental Setup

Backbones, Data, and Metrics. Our primary backbone is the Indic-specialized Sarvam-1 (2B) (Sarvam AI, 2024), with Llama-3.2 (3B) (Dubey et al., 2024) and Qwen3 (1.7B) (Yang et al., 2025) as secondary backbones, in every case with LoRA (Hu et al., 2022) adapters on eight Indic languages of Indo-Aryan and Dravidian typology (Hindi, Marathi, Bengali, Gujarati, Tamil, Telugu, Kannada, Malayalam). Training draws on 44,160 reference pairs and 57,378 rule-generated counterfactual pairs (§2.5) from released Indic instruction resources (Khan et al., 2024), and evaluation uses one held-out set of 11,978 pairs, 5,520 reference and 6,458 corrupted, across wrong case, agreement, honorificity, and word-order errors, where a pair counts as correct when the preferred response scores higher. A reference pair contrasts a released response with a word-level shuffle of that same response, hence the reference column measures sensitivity to word order rather than a human judgment between two fluent candidates. Besides *overall win-rate*, *reference-pair accuracy*, and *morphology minimal-pair accuracy*, we report two tail-sensitive quantities over language-script-morphology groups, *worst-group win-rate* and *group gap* (best minus worst, lower is better). Our primary metric is *composite accuracy*, $0.4\,$(reference-pair acc.) $+\ 0.4\,$(minimal-pair acc.) $+$ $0.2\,$(worst-group win-rate), rewarding ranking quality, morphological correctness, and tail robustness jointly, and Appendix D defines each metric.

Baselines and Training. Under one data budget and protocol per backbone we compare against DPO (Rafailov et al., 2023), SimPO (Meng et al., 2024), ORPO (Hong et al., 2024), KTO (Ethayarajh et al., 2024), Token-level DPO (TDPO) (Zeng et al., 2024), Token-level Importance Sampling DPO (TIS-DPO) (Liu et al., 2025), and the morphology-aware-tokenization variant MorphTok-DPO (Brahma et al., 2025). ORPO and KTO run on Sarvam-1 only, the other five baselines and MorphDPO on all three backbones under one shared configuration. Every method trains for one epoch of LoRA fine-tuning on a single 48 GB GPU, with hyperparameters selected by Optuna (Akiba et al., 2019) on a held-out split and frozen across runs and ablations, fixing β=0.061, λ_{PO}=2.43 and η=1.0. Morphology weights are bounded to $[1, 3]$ (§2.2), the paradigm and supervised terms of equation 10 carry zero weight in the selected trial, and MorphDPO marginalizes over up to 3 analyses per response. The margin scale of equation 6 is zero in the selected trial, hence δ_e=0 everywhere, and counterfactual pairs carry reliability γ_c=0.8 for the four substitution types and γ_c=0.7 for the word-order transposition, with no human-verified pair in the data. The search treated the tilted-policy $\log Z$ correction as a binary choice and selected the disabled setting, hence every full-scale MorphDPO number uses the morphology-weighted, length-normalized reward (α=1) without the $\log Z$ term, whose effect §3.3 isolates. Appendices E and F contain full training details and baseline objectives.

3.2 Ranking Quality and the Morphology Tax (RQ1, RQ2)

MorphDPO attains the best composite accuracy (0.9413) and overall win-rate (0.9848), answering RQ1 (Table 1). DPO is the closest competitor (0.9394), ahead on reference-pair accuracy (0.9922 vs. 0.9899) yet behind on minimal-pair accuracy (0.9732 vs. 0.9805) and tied on worst-group win-rate (0.7660) and gap (0.2340). SimPO, TDPO, and KTO pay for that ranking with group gaps of 0.32 to 0.55. Reference-pair accuracy spans only 0.9357 to 0.9922 across the eight objectives,

Table 1: Preference-ranking on the held-out morphology test set, where group gap is best-group minus worst-group win-rate (lower is better). Best per column and backbone in **bold**.

Backbone	Method	Composite Acc. ↑	Reference-Pair Acc. ↑	Morph Min-Pair Acc. ↑	Worst-Group Win-Rate ↑	Group Gap ↓	Overall Win-Rate ↑
Sarvam-1 (2B)	**Ours**	**0.9413**	0.9899	**0.9805**	**0.7660**	**0.2340**	**0.9848**
	DPO	0.9394	**0.9922**	0.9732	**0.7660**	**0.2340**	0.9820
	ORPO	0.9290	0.9909	0.9698	0.7234	0.2766	0.9795
	SimPO	0.9038	0.9889	0.9300	0.6809	0.3191	0.9572
	TDPO	0.8951	0.9888	0.9299	0.6383	0.3592	0.9570
	KTO	0.8099	0.9889	0.8092	0.4531	0.5469	0.8921
	TIS-DPO	0.8098	0.9357	0.8548	0.4681	0.5191	0.8921
	MorphTok-DPO	0.7483	0.9899	0.6891	0.3835	0.6165	0.8277
Llama-3.2 (3B)	**Ours**	**0.8118**	0.9654	**0.8726**	**0.3830**	**0.6170**	**0.9153**
	DPO	0.7689	**0.9674**	0.8167	0.2766	0.7220	0.8861
	SimPO	0.7765	0.9627	0.8295	0.2979	0.7002	0.8909
	TDPO	0.7789	0.9652	0.8117	0.3404	0.6577	0.8825
	TIS-DPO	0.7050	0.9567	0.6994	0.2128	0.7830	0.8180
	MorphTok-DPO	0.7389	0.9663	0.8066	0.1489	0.8496	0.8802
Qwen3 (1.7B)	**Ours**	**0.8387**	0.9678	**0.8950**	**0.4681**	**0.5304**	**0.9285**
	DPO	0.8079	**0.9697**	0.8692	0.3617	0.6359	0.9155
	SimPO	0.8085	0.9690	0.8501	0.4043	0.5934	0.9049
	TDPO	0.8103	0.9687	0.8444	0.4255	0.5721	0.9017
	TIS-DPO	0.7245	0.9513	0.7004	0.3191	0.6733	0.8160
	MorphTok-DPO	0.7676	0.9694	0.8219	0.2553	0.7418	0.8899

while composite accuracy spans 0.7483 to 0.9413, hence the separation between methods rests on morphology rather than on the reference pairs. Figure 2 demonstrates the trade-off explicitly.

The lower two blocks of Table 1 repeat the comparison on Llama-3.2 and Qwen3 for the six shared objectives. Absolute scores drop on backbones with less Indic pretraining, yet the top of the ranking holds. MorphDPO leads on composite accuracy (0.8118 and 0.8387 against 0.7789 and 0.8103 for the best baseline), minimal-pair accuracy, worst-group win-rate, and every corruption type, while DPO keeps its reference-pair edge (Figure 7 in Appendix G).

Table 3 in Appendix G breaks accuracy down by corruption type, answering RQ2. MorphDPO is strongest on the three agreement- and case-sensitive categories, wrong case (0.9865), gender (0.9779) and number agreement (0.9372), and within 1 point of the best on honorificity and word order. No baseline matches this profile, since the baselines that rank reference pairs competitively lose as much as 42 points on these categories, leaving no morphology tax. Across backbones, MorphDPO is best in every category on Llama-3.2 and Qwen3, with the widest margins on the agreement categories, where it leads the next method by 4 to 16 points (Figure 3). Number and gender agreement stay the hardest categories there, since our accuracy of 0.66 to 0.73 trails our own word-order accuracy by 23 to 29 points, against under 5 points on Sarvam-1. TIS-DPO is weakest on case, gender, and number agreement across those backbones at 0.31 to 0.46, and MorphTok-DPO combines word-order accuracy above 0.94 with number agreement below 0.48. Figure 5 depicts the per-group win-rate distribution.

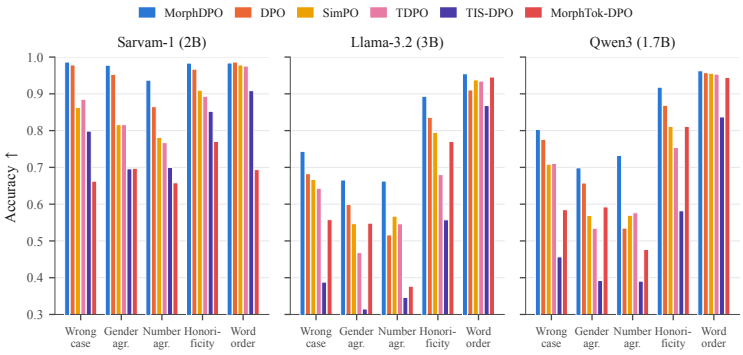


Figure 3: Accuracy by corruption type on each backbone for the six shared objectives.

Per-language behavior. Figure 4 breaks the overall win-rate down by language. On Sarvam-1 MorphDPO stays above 0.95 on all eight languages and plain DPO covers the same band, hence per-language means alone do not separate the two strongest methods there, and the separation lives in the finer language-script-morphology groups of §2.6. On both general-purpose backbones every method is strongest on Hindi and, with the single exception of TIS-DPO on Qwen3, weakest on Telugu, while MorphDPO holds the highest or joint-highest win-rate in six of the eight languages on Llama-3.2 and seven of the eight on Qwen3. The per-language spread also widens off the Indic-

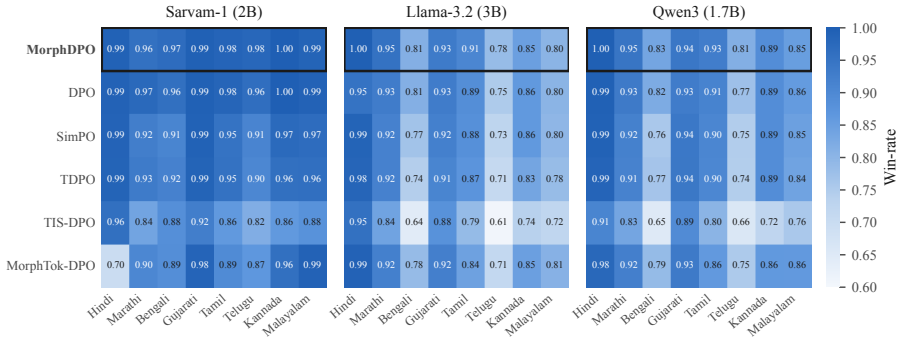


Figure 4: Per-language win-rate of the six shared objectives on each backbone.

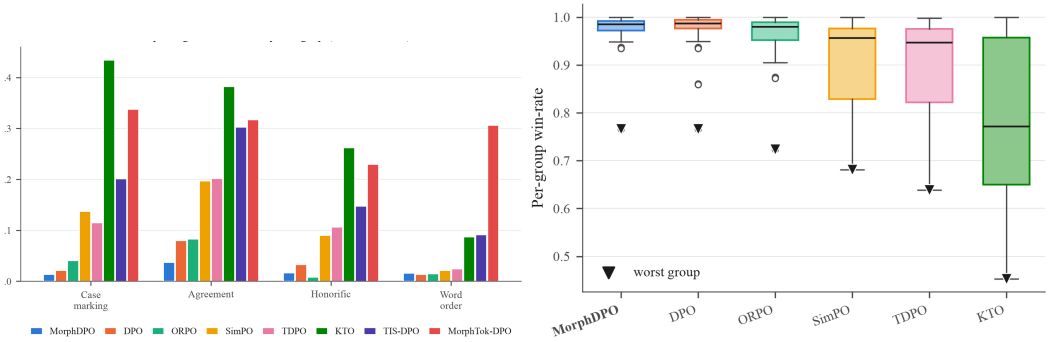


Figure 5: Tail behavior of MorphDPO and the baselines. Left: morphological error rate by category (lower is better). Right: per-group win-rate distribution with worst-group markers (▼), where MorphDPO has the tightest spread and shares the top worst group with DPO.

specialized backbone, from 4 points between the best and the worst language on Sarvam-1 to 21 on Llama-3.2 and 18 on Qwen3. Telugu stays the hardest language on both general-purpose backbones, where MorphDPO still leads the strongest baseline by 3.3 and 3.9 points, and the two languages left unled are Kannada on Llama-3.2 and Malayalam on both general-purpose backbones. On Sarvam-1 the weakest language for MorphDPO is instead Marathi at 0.956, with Telugu at 0.979.

3.3 Component Ablations (RQ3)

Subset runs vary one factor at a time on the same test set (Table 2, and Figure 9 in Appendix H). Each block is an independent set of runs, hence the chosen variant carries a different score in each block, and the posterior block marginalizes over up to 5 analyses rather than the 3 of the frozen configuration. Two of these runs share a configuration exactly, namely the no-correction arm of the $\log Z$ block and the normalized arm of the reward block, both of which use the normalized reward without the $\log Z$ term under seed 42. The pair lands at 0.8694 and 0.8892 composite accuracy and at 0.5957 and 0.6809 worst-group win-rate, which puts the run-to-run spread of these subset runs at 2.0 and 8.5 points on the two metrics and sets the bar that any single-factor effect has to clear.

Read against that spread, one effect on Sarvam-1 stands clear. The length-and-density normalization (α=1 in equation 5) lifts worst-group win-rate from 0.5532 to 0.6809 and shrinks the gap from 0.4402 to 0.3121, because an unnormalized weighted sum over-rewards long, morphology-dense responses, and the unnormalized variant scores higher on reference-pair and minimal-pair accuracy, trading a little average ranking for a stronger tail. The other two effects fall inside the spread. Marginalizing over the analyzer posterior (§2.1) rather than a single 1-best analysis moves composite accuracy from 0.8746 to 0.8794 and worst-group win-rate from 0.6170 to 0.6364, and the exact tilted-policy $\log Z$ correction (§2.3) reaches 0.6383 worst-group against 0.5957 without the correction and 0.5745 with an approximate one, differences of 2 to 6 points on a metric whose replicate spread is 8.5. The frozen configuration switches the correction *off* in any case (§3.1). None of the three effects replicates on Llama-3.2 and Qwen3, where variants within each block land within about two points of one another and the unnormalized reward scores higher on composite accuracy (Appendix G), which attributes the cross-backbone advantage to the objective as a whole. Appendix C connects these choices to the results of §3.2.

Table 2: Component ablations on the held-out morphology test set, one factor varied per block over independent subset runs. Best per block in **bold**.

Factor	Variant	Composite Acc. ↑	Reference-Pair Acc. ↑	Morph Min-Pair Acc. ↑	Worst-Group Win-Rate ↑	Group Gap ↓
Analysis posterior	Hard 1-best	0.8746	0.9739	0.9041	0.6170	0.3764
	Posterior (ours)	**0.8794**	**0.9748**	**0.9055**	**0.6364**	**0.3565**
$\log Z$ normalization	None	0.8694	**0.9739**	0.9018	0.5957	0.3972
	Approximate	0.8659	0.9728	**0.9046**	0.5745	0.4189
	Exact (ours)	**0.8787**	0.9734	0.9043	**0.6383**	**0.3551**
Reward normalization	Unnormalized	0.8675	**0.9793**	**0.9128**	0.5532	0.4402
	Normalized (ours)	**0.8892**	0.9736	0.9091	**0.6809**	**0.3121**

4 Related Work

Preference Optimization Objectives. Reinforcement learning from human feedback (RLHF) fits a reward model to pairwise comparisons and optimizes the policy against it (Christiano et al., 2017; Ouyang et al., 2022), and DPO folds both stages into one logistic loss on the sequence log-ratio (Rafailov et al., 2023). Later objectives drop the frozen reference (SimPO, ORPO) (Meng et al., 2024; Hong et al., 2024), learn from unpaired responses (KTO) (Ethayarajh et al., 2024), or guard against label noise (Dr. DPO) (Wu et al., 2025). Token-level variants sharpen the training signal with a per-token Kullback-Leibler regularizer (TDPO) (Zeng et al., 2024), estimated token importance (TIS-DPO) (Liu et al., 2025), or self-generated reward regularization (T-REG) (Zhou et al., 2025). All define the implicit reward over whole sequences or subword tokens, and the reweighting variants leave the weighted sum of log-ratios unnormalized. MorphDPO instead defines the reward over a latent morpheme lattice, each weighted term the log-ratio of a normalized tilted distribution (§2.3) that recovers DPO exactly under uniform weights, and our reliability-weighted loss (§2.4) discounts each synthetic morphology label by a per-edge reliability.

Morphology, Robustness, and Indic Alignment. Morphological feature bundles follow UniMorph (Batsuren et al., 2022), and group-DRO addresses subpopulation shift (Sagawa et al., 2020), applied here over language-script-morphology groups rather than noisy pairs (§2.6). Prior Indic and multilingual work improves inputs, data, or the reward model rather than the objective, through morphology-grounded tokenization with sandhi splitting (MorphTok) (Brahma et al., 2025), reasoning alignment from weaker languages to a dominant one (MAPO) (She et al., 2024), and retrieval that extends reward models to low-resource Indic languages (RELIC) (Ghosal et al., 2025), all built on Indic-specific resources (Khan et al., 2024; Sarvam AI, 2024; Singh et al., 2024). Morphology-aware tokenization alone does not remove the morphology tax (MorphTok-DPO in Table 1), and the objective has to operate on morphological units.

5 Conclusion

Alignment in morphologically rich languages runs into a mismatch of granularity. Localized grammatical units decide the preference, while standard objectives score whole sequences or compression-optimized subword tokens. **MorphDPO** closes this gap by moving the unit of alignment to a latent morpheme lattice. Its reward is a normalized tilted-policy log-ratio that is tokenizer-agnostic, marginalizes over uncertain analyses, and reduces exactly to DPO under uniform weights, and around that reward we build counterfactual partial-order preferences, a paradigm-neighborhood robust objective, and a group-distributionally-robust objective, each with a guarantee (Appendix B). On the Indic-specialized Sarvam-1 (2B), MorphDPO attains the best composite accuracy (0.941) and overall win-rate (0.985) among preference objectives and a group gap tied with DPO for the smallest, making it the only preference-tuned method that improves ranking without a "morphology tax," and the same lead holds on Llama-3.2 (3B) and Qwen3 (1.7B). Therefore, we propose the morpheme as the general unit of alignment for languages whose acceptability is governed by structure below the word.

Limitations. Our experiments cover three backbones of at most 3B parameters and eight Indic languages, and the method depends on a morphological analyzer, with token-level behavior as the fallback wherever none is reliable. Our evaluation also inherits its data construction, since the corrupted pairs come from a rule catalogue rather than from a certified validity predicate, the largest corrupted category tests word order rather than a morphological slot, and the reference pairs contrast a released response with a shuffle of itself.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. UniMorph 4.0: Universal Morphology. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 840–855, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.89/.

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.

Maharaj Brahma, N J Karthika, Atul Kumar Singh, Devaraja Adiga, Smruti Bhate, Ganesh Ramakrishnan, Rohit Saluja, and Maunendra Sankar Desarkar. Morphtok: Morphologically grounded tokenization for indic languages. In *ICML 2025 Tokenization Workshop (TokShop)*, 2025. URL https://openreview.net/forum?id=32wHtvZOqH.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pp. 4299–4307. Curran Associates, Inc., 2017.

Abhimanyu Dubey, Akhil Jauhri, Ankit Pandey, Abhishek Kadian, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

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.

Soumya Suvra Ghosal, Vaibhav Singh, Akash Ghosh, Soumyabrata Pal, Subhadip Baidya, Sriparna Saha, and Dinesh Manocha. RELIC: Enhancing reward model generalization for low-resource Indic languages with few-shot examples. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1502–1517, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.79. URL https://aclanthology.org/2025.findings-emnlp.79/.

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.

Mohammed Safi Ur Rahman Khan, Priyam Mehta, Ananth Sankar, Umashankar Kumaravelan, Sumanth Doddapaneni, Suriyaprasaad B, Varun G, Sparsh Jain, Anoop Kunchukuttan, Pratyush Kumar, Raj Dabre, and Mitesh M. Khapra. IndicLLMSuite: A blueprint for creating pre-training and fine-tuning datasets for Indian languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15831–15879, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.843. URL https://aclanthology.org/2024.acl-long.843/.

Aiwei Liu, Haoping Bai, Zhiyun Lu, Yanchao Sun, Xiang Kong, Xiaoming Simon Wang, Jiulong Shan, Albin Madappally Jose, Xiaojiang Liu, Lijie Wen, Philip S. Yu, and Meng Cao. TIS-DPO: Token-level importance sampling for direct preference optimization with estimated weights. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=oF6e2WwxX0.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37, pp. 124198–124235, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/e099c1c9699814af0be873a175361713-Abstract-Conference.html.

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.

Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pp. 745–750. ACM, 2007.

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*, volume 36, pp. 53728–53742, 2023.

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.

Sarvam AI. Sarvam 1: The first indian language llm. https://www.sarvam.ai/blogs/sarvam-1, October 2024. Accessed: 2026-09-25.

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

Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11047–11073, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.595. URL https://aclanthology.org/2024.acl-long.595/.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020. doi: 10.1162/tacl_a_00321. URL https://aclanthology.org/2020.tacl-1.25/.

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

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.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60315–60334. PMLR, 2024. URL https://proceedings.mlr.press/v235/zeng24c.html.

Wenxuan Zhou, Shujian Zhang, Lingxiao Zhao, and Tao Meng. T-REG: Preference optimization with token-level reward regularization. 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. 27876–27889, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1353. URL https://aclanthology.org/2025.acl-long.1353/.

SUPPLEMENTARY MATERIAL

A Method Details **13**

B Theoretical Guarantees and Proofs **14**

B.1 DPO Reduction 14

B.2 Analyzer-Uncertainty Bound 14

B.3 Counterfactual Correctness 15

B.4 Paradigm and Worst-Group Control 15

C Why Morphology-Consistent Optimization Helps **16**

D Metric Definitions **17**

E Detailed Experimental Setup **18**

F Baseline Objectives **19**

G Full Results **20**

H Training Dynamics and Per-Language Results **23**

H.1 Per-Language Behavior 23

I Analyses **24**

I.1 Qualitative Analysis 24

I.2 Error Analysis 25

J License and Assets **26**

K Broader Impact **26**

L AI Use Statement **26**

A METHOD DETAILS

Frequency-Aware Morphology Weights: Beyond the linguistic prior of §2.2, a frequency-aware variant sets $w(\phi, \ell) = \text{clip}\left(1 + \zeta\, s(\phi) + \mu\,(\hat{f}_{\ell,\phi} + \epsilon)^{-1/2},\ 1,\ W\right)$, where $s(\phi) \in [0, 1]$ is a fixed severity score, $\hat{f}_{\ell,\phi}$ is the empirical frequency of feature ϕ in language ℓ, and $\zeta, \mu \geq 0$ are tuned on the validation split. The clipping to $[1, W]$ retains the guarantees of §2.3, since those depend only on the weights being bounded.

Evaluating the Tilted Normalizer: Because the weights of §2.3 are computed from an analysis of y, the guarantees in Appendix B require only the per-position normalization. The normalizer itself is evaluated as

$$\log Z_\theta(h_t, \omega) = \operatorname*{logsumexp}_{v \in \mathcal{V}} \Big[\log \pi_0(v \mid h_t) + \omega\big(\log \pi_\theta(v \mid h_t) - \log \pi_0(v \mid h_t)\big)\Big], \tag{11}$$

which adds a full-vocabulary normalization at target positions without differentiating through the analyzer.

Training Pipeline: For each response in a preference edge the analyzer produces candidate analyses and feature bundles, character-span overlaps induce token-level morphology weights equation 2,

the policy and reference produce token logits, and the reward follows from the normalized tilted log-ratio equation 4. Counterfactual edges are generated offline and filtered by equation 7, while paradigm neighborhoods are generated offline or sampled during training. Without a usable segmentation, all positions take $\omega_t = 1$ and the objective reduces locally to DPO, or to length-normalized DPO depending on α, which leaves the objective well defined for languages with incomplete morphology tools.

Supervised Regularizer: The optional term in equation 10 is $\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{sup}}}\big[\frac{1}{|y|}\sum_t \log \pi_\theta(u_t \mid x, u_{<t})\big]$, applied with a small coefficient to preserve fluency and instruction-following ability.

B Theoretical Guarantees and Proofs

This appendix states and proves the formal guarantees of MorphDPO referenced at the end of §2. Throughout, σ denotes the logistic function, $h_t = (x, u_{<t})$, and all sums over the vocabulary $\mathcal{V}$ are finite, which leaves exchanges of summation and the existence of the normalizer Z_θ unproblematic.

B.1 DPO Reduction

MorphDPO first strictly generalizes DPO.

Proposition 1 (DPO reduction). *If every analysis is deterministic ($q_\ell(a \mid y) = 1$) and all morphology weights are uniform ($\omega_t(a) = 1$ for all t), then $Z_\theta(h_t, 1) = 1$ and, with $\alpha = 0$, $R^M_\theta(x, y, \ell) = \log \frac{\pi_\theta(y|x)}{\pi_0(y|x)}$. Consequently, if in addition every edge is a fully trusted reference pair ($\gamma_e = 1$, $\delta_e = 0$), all edges share a single group, there are no counterfactual or paradigm edges ($\lambda_{\mathrm{PO}} = \lambda_{\mathrm{par}} = 0$), and $\lambda_{\mathrm{SFT}} = 0$, then $\mathcal{J}_{\mathrm{MorphDPO}}$ equals the DPO objective equation 1 averaged over the data.*

Proof. Fix $\alpha = 0$ and suppose $q_\ell(\cdot \mid y)$ is a point mass on a single analysis a with $\omega_t(a) = 1$ for all t. Evaluating the normalizer equation 3 at $\omega = 1$,

$$Z_\theta(h_t, 1) = \sum_{v\in\mathcal{V}} \pi_\theta(v \mid h_t)^1 \, \pi_0(v \mid h_t)^0 = \sum_{v\in\mathcal{V}} \pi_\theta(v \mid h_t) = 1, \tag{12}$$

and therefore $\log Z_\theta(h_t, 1) = 0$. The per-token contribution equation 4 then collapses to the plain log-ratio, $\rho^M_\theta(u_t, h_t, a) = \log \frac{\pi_\theta(u_t|h_t)}{\pi_0(u_t|h_t)}$. With $\alpha = 0$ the denominator of equation 5 equals $\big(\sum_t \omega_t(a)\big)^0 = 1$, and since q_ℓ is degenerate the expectation in equation 5 is redundant, giving

$$R^M_\theta(x, y, \ell) = \sum_{t=1}^{T} \log \frac{\pi_\theta(u_t \mid h_t)}{\pi_0(u_t \mid h_t)} = \log \frac{\pi_\theta(y \mid x)}{\pi_0(y \mid x)} = r_\theta(x, y). \tag{13}$$

Consequently the margin becomes $m_\theta(e) = r_\theta(x, y^+) - r_\theta(x, y^-)$, and for a fully trusted reference pair ($\gamma_e = 1$, $\delta_e = 0$) the reliability-weighted loss equation 6 reduces to $\ell(e; \theta) = -\log \sigma(\beta[r_\theta(x, y^+) - r_\theta(x, y^-)])$, which is exactly $\mathcal{L}_{\mathrm{DPO}}$ in equation 1. Under the remaining hypotheses, $\mathcal{B}$ contains a single group b with no parent, and the shrinkage target is the global mean, which coincides with $\bar{\mathcal{L}}_b$. Hence $\widetilde{\mathcal{L}}_b = \bar{\mathcal{L}}_b$ and $\mathcal{L}_{\mathrm{DRO}} = \eta \log \exp(\bar{\mathcal{L}}_b/\eta) = \bar{\mathcal{L}}_b$, the mean of the per-edge DPO losses. With $\lambda_{\mathrm{PO}} = \lambda_{\mathrm{par}} = \lambda_{\mathrm{SFT}} = 0$ the remaining terms of equation 10 vanish, and $\mathcal{J}_{\mathrm{MorphDPO}}$ equals the DPO objective averaged over the data. □

B.2 Analyzer-Uncertainty Bound

Analyzer uncertainty perturbs the objective in a controlled way, addressing the natural concern that morphological analyzers are imperfect.

Theorem 1 (Analyzer-uncertainty bound). *Let $p(a \mid y)$ be the true posterior over analyses and $q(a \mid y)$ the analyzer posterior, with $\mathrm{TV}(p,q) \le \varepsilon$ and $|R_\theta^M(x,y,a)| \le B$. Then $\left|\mathbb{E}_p R_\theta^M(x,y,a) - \mathbb{E}_q R_\theta^M(x,y,a)\right| \le 2B\varepsilon$. For a preference pair the margin error is at most $4B\varepsilon$, and since $-\log\sigma(\beta m)$ is β-Lipschitz in m, the loss perturbation is at most $4\beta B\varepsilon$.*

Proof. Set $f(a) := R_\theta^M(x,y,a)$, and recall the identity relating total variation to the ℓ_1 distance for distributions over a countable set, $\sum_a |p(a \mid y) - q(a \mid y)| = 2\ \mathrm{TV}(p,q)$. Using $|f(a)| \le B$,

$$|\mathbb{E}_p f - \mathbb{E}_q f| = \left|\sum_a \big(p(a \mid y) - q(a \mid y)\big) f(a)\right| \le \sum_a \big|p(a \mid y) - q(a \mid y)\big|\,|f(a)|$$
$$\le B \sum_a \big|p(a \mid y) - q(a \mid y)\big| = 2B\ \mathrm{TV}(p,q) \le 2B\varepsilon. \tag{14}$$

Since $R_\theta^M(x,y,\ell) = \mathbb{E}_a f(a)$ by equation 5, applying this bound to both responses in a preference pair and using the triangle inequality, the margin $m_\theta(e) = R_\theta^M(x,y^+,\ell) - R_\theta^M(x,y^-,\ell)$ changes by at most $2B\varepsilon + 2B\varepsilon = 4B\varepsilon$ when q replaces p. Finally, the map $m \mapsto -\log\sigma(\beta(m-\delta))$ has derivative $-\beta\,(1-\sigma(\beta(m-\delta)))$, whose magnitude is bounded by β, hence β-Lipschitz in m. The reliability-weighted loss equation 6 is a convex combination of two such β-Lipschitz maps (in m and $-m$) and is therefore also β-Lipschitz in m. Composing the Lipschitz bound with the margin bound yields a loss perturbation of at most $\beta \cdot 4B\varepsilon = 4\beta B\varepsilon$. □

B.3 Counterfactual Correctness

The counterfactual construction is justified under a mild utility decomposition.

Theorem 2 (Counterfactual correctness). *Suppose human utility decomposes as $U^*(x,y) = U_{\mathrm{sem}}(x,y) - \sum_c \kappa_c E_c(y)$ with $\kappa_c > 0$ and $E_c(y) \in \{0,1\}$ indicating morphology error type c (the utility penalty κ_c is distinct from the label reliability γ_c of §2.5). If y is free of the target error, $E_c(y) = 0$, and T_c preserves semantics, $U_{\mathrm{sem}}(x,y) = U_{\mathrm{sem}}(x,T_c(y))$, and introduces exactly the target error, $E_c(T_c(y)) = 1$ with all other indicators unchanged, then $U^*(x,y) - U^*(x,T_c(y)) = \kappa_c > 0$, hence $y \succ T_c(y)$. The validity predicate equation 7 states precisely these assumptions, and §3.1 reports how far our generator enforces them.*

Proof. By the assumed utility decomposition,

$$U^*(x,y) - U^*(x,T_c(y)) = \underbrace{\big[U_{\mathrm{sem}}(x,y) - U_{\mathrm{sem}}(x,T_c(y))\big]}_{=0} - \sum_{c'} \kappa_{c'} \big[E_{c'}(y) - E_{c'}(T_c(y))\big]. \tag{15}$$

The first bracket vanishes by semantic preservation. For $c' \neq c$ the indicators are unchanged, and $E_{c'}(y) - E_{c'}(T_c(y)) = 0$. For $c' = c$ the transformation introduces exactly the target error into an error-free response, $E_c(y) = 0$ and $E_c(T_c(y)) = 1$, giving $E_c(y) - E_c(T_c(y)) = -1$. Therefore

$$U^*(x,y) - U^*(x,T_c(y)) = -\kappa_c \cdot (-1) = \kappa_c > 0. \tag{16}$$

Under the Bradley-Terry model (Bradley & Terry, 1952), $P^*(y \succ T_c(y) \mid x) = \sigma\big(\beta^*[U^*(x,y) - U^*(x,T_c(y))]\big) = \sigma(\beta^*\kappa_c) > \frac{1}{2}$, and y is preferred to $T_c(y)$, which makes the counterfactual label $y \succ T_c(y)$ correct. The validity predicate equation 7 accepts a pair only when $\mathrm{SemPres}$, $\mathrm{TargetErr}_c$, and $\bigwedge_{c' \neq c} \mathrm{NoExtraErr}_{c'}$ all hold, which are precisely the three hypotheses used above ($\mathrm{TargetErr}_c$ is evaluated on a response y that passed the same check, giving $E_c(y) = 0$). The margin δ_c may be set to κ_c, or to a lower bound thereof, to reflect the utility gap. □

B.4 Paradigm and Worst-Group Control

Finally, the paradigm and group objectives are smooth upper bounds on the corresponding worst-case risks.

Theorem 3 (Paradigm and worst-group control)**.** *For any edge e, $\mathbb{E}_{g\sim\mathcal{G}_\ell(e)}[\ell(g(e);\theta)] \le \max_g \ell(g(e);\theta) \le \mathcal{L}_{\mathrm{par}}(e)$. For the group objective, $\max_b \widetilde{\mathcal{L}}_b(\theta) \le \mathcal{L}_{\mathrm{DRO}}(\theta) \le \max_b \widetilde{\mathcal{L}}_b(\theta) + \eta \log|\mathcal{B}|$. Moreover, because all per-edge losses are non-negative, the unshrunk group losses obey $\max_b \bar{\mathcal{L}}_b(\theta) \le \left(\max_b \frac{n_b+\rho}{n_b}\right) \mathcal{L}_{\mathrm{DRO}}(\theta)$. Thus minimizing equation 8 and equation 9 controls worst-transformation and worst-group morphology loss, respectively.*

Proof. Both claims follow from the elementary sandwich bounds for log-sum-exp. Let $\{z_i\}_{i=1}^n$ be real numbers and $\kappa > 0$. Since $\exp(\max_i z_i/\kappa) \le \sum_i \exp(z_i/\kappa) \le n \exp(\max_i z_i/\kappa)$, taking $\kappa \log(\cdot)$ (monotone) yields

$$\max_i z_i \;\le\; \kappa \log \sum_{i=1}^{n} \exp\Big(\frac{z_i}{\kappa}\Big) \;\le\; \max_i z_i + \kappa \log n. \tag{17}$$

Paradigm neighborhood: Apply the left inequality of equation 17 with $z_g = \ell(g(e);\theta)$, $\kappa = \tau_g$, and $n = |\mathcal{G}_\ell(e)|$, which yields $\max_g \ell(g(e);\theta) \le \tau_g \log \sum_g \exp(\ell(g(e);\theta)/\tau_g) = \mathcal{L}_{\mathrm{par}}(e)$ from equation 8. Because a mean cannot exceed a maximum, $\mathbb{E}_{g\sim\mathcal{G}_\ell(e)}[\ell(g(e);\theta)] \le \max_g \ell(g(e);\theta)$, and any distribution over valid paradigm transformations, in particular the test distribution of §2.6, has expected loss upper-bounded by $\mathcal{L}_{\mathrm{par}}(e)$. The right inequality of equation 17 additionally yields $\mathcal{L}_{\mathrm{par}}(e) \le \max_g \ell(g(e);\theta) + \tau_g \log|\mathcal{G}_\ell(e)|$, and the surrogate is tight up to the temperature term, recovering the exact worst case as $\tau_g \to 0^+$.

Worst group: Apply equation 17 with $z_b = \widetilde{\mathcal{L}}_b(\theta)$, $\kappa = \eta$, and $n = |\mathcal{B}|$,

$$\max_{b\in\mathcal{B}} \widetilde{\mathcal{L}}_b(\theta) \;\le\; \eta \log \sum_{b\in\mathcal{B}} \exp\left(\frac{\widetilde{\mathcal{L}}_b(\theta)}{\eta}\right) = \mathcal{L}_{\mathrm{DRO}}(\theta) \;\le\; \max_{b\in\mathcal{B}} \widetilde{\mathcal{L}}_b(\theta) + \eta \log|\mathcal{B}|, \tag{18}$$

which is exactly the stated sandwich for equation 9. Hence minimizing $\mathcal{L}_{\mathrm{DRO}}$ minimizes an upper bound on the worst-group (shrinkage-adjusted) loss, and the bound is loose by at most the known additive constant $\eta \log|\mathcal{B}|$. Finally, the per-edge loss equation 6 is a convex combination of negative log-sigmoids and is therefore non-negative, and every group mean $\bar{\mathcal{L}}_b$ and parent mean $\bar{\mathcal{L}}_{\mathrm{pa}(b)}$ is non-negative. Dropping the parent term in the shrinkage estimator yields $\widetilde{\mathcal{L}}_b \ge \frac{n_b}{n_b+\rho} \bar{\mathcal{L}}_b$, i.e. $\bar{\mathcal{L}}_b \le \frac{n_b+\rho}{n_b} \widetilde{\mathcal{L}}_b \le \frac{n_b+\rho}{n_b} \mathcal{L}_{\mathrm{DRO}}$, and taking the maximum over b yields the stated control of the unshrunk worst-group loss. □

C Why Morphology-Consistent Optimization Helps

Below we describe the mechanistic reasons that connect MorphDPO's design to its empirical behavior in §3. The account is intended to be falsifiable, in that each claim points to a specific measurement.

(i) The unit of the training signal matches the unit of failure. In morphologically rich languages a response is often rejected because of a single grammatical unit, such as a case marker or an agreement or TAM suffix, rather than because of the response as a whole. MorphDPO places the preference gradient on these units by weighting the reward over a morpheme lattice, instead of spreading the gradient across compression-optimized subword tokens. The prediction is that the method's advantage should be largest exactly on the localized suffixal contrasts and smallest where the distinguishing signal is already dense. This is what Table 3 reports, with accuracies of 0.987 on wrong case, 0.978 on gender agreement, and 0.937 on number agreement, up to tens of points above the reference-free and token-level baselines, and with the widest margin over the runner-up on number agreement. On word order, the largest and most surface-frequent category, every method on Sarvam-1 except MorphTok-DPO lands above 0.90.

(ii) Normalization is what protects the tail. Two normalizations enter the reward. The length-and-density normalization of equation 5 (α=1) divides the weighted sum of token log-ratios by the total morphology weight, which stops long or morphology-dense responses from being rewarded for

their length. The tilted-policy $\log Z$ correction of equation 4 keeps each weighted token term the log-ratio of a normalized distribution. In the controlled ablation (Table 2, Figure 9) the first is the single largest effect, since on identical data the normalized reward raises worst-group win-rate from 0.553 to 0.681 and shrinks the group gap from 0.440 to 0.312. The exact $\log Z$ correction adds a smaller gain over no correction (0.638 vs. 0.596 worst-group) that falls inside the replicate spread of §3.3 and therefore carries no weight on its own. Because the degradation without normalization concentrates on the rarest groups, this is direct evidence that the objective's probabilistic structure matters for the tail. The frozen full-scale configuration uses the length normalization without the $\log Z$ correction (§3.1).

(iii) Why competing preference objectives pay a "morphology tax." Reference-free, token-level, and unpaired objectives (SimPO, TDPO, KTO) raise reference-pair ranking by moving probability mass toward globally fluent-looking completions. Fluency correlates only weakly with morphological correctness, and their advantage therefore does not transfer to corrupted minimal pairs, while their average-to-worst-group gaps widen to 0.32 to 0.55 (Table 1). MorphDPO avoids this because its gradient is anchored to morpheme-level contrasts and normalized per position, which improves ranking to the best composite accuracy (0.941) and overall win-rate (0.985) while staying within about a point of the strongest per-corruption accuracies. Figure 2 visualizes this as the only point that is top-right in both the ranking-morphology and overall-worst-group planes. The fluency-versus-morphology account above is our interpretation of the correlation, and the measured facts are the joint ranking-and-morphology position and the gap values.

(iv) Objective versus data. Because MorphDPO also trains on counterfactual minimal pairs, part of its gain could in principle stem from the added data rather than the objective. What the controlled ablations attribute to the objective is within-objective tail robustness, since the normalization ablation in (ii) holds data fixed and still moves the worst group by more than the replicate spread of §3.3, which the analyzer-marginalization and $\log Z$ comparisons do not. Separating the remaining contribution of the counterfactual pairs from that of the lattice reward at full scale, in particular with the exact $\log Z$ correction enabled, is the most important follow-up experiment.

(v) Robustness to imperfect morphology. Analyzers are noisy, and MorphDPO therefore marginalizes over the analysis posterior rather than committing to a single segmentation. Empirically the marginalization moves composite, minimal-pair, and worst-group metrics together over a hard 1-best analysis (Table 2), by margins inside the replicate spread of §3.3. Theoretically the marginalization is safe, because Theorem 1 bounds the reward perturbation by $2B\varepsilon$ and the loss perturbation by $4\beta B\varepsilon$ in the analyzer's total-variation error ε, and segmentation ambiguity therefore degrades the objective gracefully rather than injecting hard errors. Together, (i) to (v) explain why the method is simultaneously the best average ranker and the most tail-robust, without a morphology tax.

D Metric Definitions

Let $\mathcal{T} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^{M}$ be the held-out set of $M = 11{,}978$ preference pairs and let $R(x, y)$ be the reward a method assigns to response y under prompt x. For DPO and its variants R is the implicit sequence log-ratio (Rafailov et al., 2023), and for MorphDPO the reward is the morpheme-lattice reward R_θ^M of equation 5. Under the Bradley-Terry model of pairwise comparison (Bradley & Terry, 1952) the probability of the observed order is $\sigma(R(x, y^+) - R(x, y^-))$, and we count a pair as correct exactly when that log-odds is positive. The win-rate on any subset $\mathcal{S} \subseteq \mathcal{T}$ is

$$\mathrm{WR}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{(x, y^+, y^-) \in \mathcal{S}} \mathbf{1}\big\{R(x, y^+) > R(x, y^-)\big\}\,. \tag{19}$$

Ranking Metrics. The *overall win-rate* is $\mathrm{WR}(\mathcal{T})$. Writing $\mathcal{T}_h$ for the $5{,}520$ reference pairs and $\mathcal{T}_m$ for the $6{,}458$ morphology-corrupted pairs, with $\mathcal{T} = \mathcal{T}_h \cup \mathcal{T}_m$, the *reference-pair accuracy* is $\mathrm{WR}(\mathcal{T}_h)$ and the *morphology minimal-pair accuracy* is $\mathrm{WR}(\mathcal{T}_m)$. The per-corruption-type accuracies of Table 3 are WR restricted to the pairs carrying each corruption type c.

Tail Metrics. Every pair belongs to one language-script-morphology group $b \in \mathcal{B}$ (§2.6), giving the partition $\mathcal{T} = \bigcup_{b\in\mathcal{B}} \mathcal{T}_b$. The *worst-group win-rate* and the *group gap* are

$$\mathrm{WG} = \min_{b\in\mathcal{B}} \mathrm{WR}(\mathcal{T}_b), \qquad \Delta = \max_{b\in\mathcal{B}} \mathrm{WR}(\mathcal{T}_b) - \min_{b\in\mathcal{B}} \mathrm{WR}(\mathcal{T}_b), \tag{20}$$

with lower Δ better. WG is the empirical counterpart of the worst-group risk that group-DRO minimizes (Sagawa et al., 2020), which is what aligns the training objective of §2.6 with this evaluation.

Composite Accuracy. Our primary metric combines ranking quality, morphological correctness, and tail robustness with fixed weights,

$$C = 0.4\,\mathrm{WR}(\mathcal{T}_h) + 0.4\,\mathrm{WR}(\mathcal{T}_m) + 0.2\,\mathrm{WG}. \tag{21}$$

The weights are fixed a priori and stay untuned. Equal mass on the two accuracy terms keeps a method from buying ranking quality with morphological errors, and the remaining mass on WG prevents a strong average from hiding a failing group.

E Detailed Experimental Setup

Languages and data. We use eight Indic languages spanning Indo-Aryan (Hindi, Marathi, Bengali, Gujarati) and Dravidian (Tamil, Telugu, Kannada, Malayalam) typologies and multiple scripts. Reference pairs are split $44{,}160$/$5{,}520$/$5{,}520$ (train/validation/test). The validation split is reference pairs only and is used for checkpoint selection, while the test set adds $6{,}458$ morphology-corrupted pairs to the $5{,}520$ held-out reference pairs, giving $11{,}978$ evaluation pairs. The dispreferred side of a reference pair is a word-level shuffle of the preferred response, drawn under a seed that differs per backbone, hence the reference column compares verdicts on different negatives across backbones while the corrupted pairs stay byte-identical throughout. Corrupted pairs cover wrong case (889), gender agreement (724), number agreement (430), honorificity (122), and word order ($4{,}293$). The first four types come from a fixed catalogue of regular-expression substitutions per language, which reuses the Hindi rules for Gujarati, the Telugu rules for Kannada, and the Tamil rules for Malayalam, while the word-order type transposes the final two words of the response and applies to every language alike. A separate set of $57{,}378$ counterfactual pairs is available for methods that use counterfactual data. Morphological feature bundles follow the UniMorph schema (Batsuren et al., 2022).

Optimization and hardware. We fine-tune each backbone, namely Sarvam-1 (2B) (Sarvam AI, 2024), Llama-3.2 (3B) (Dubey et al., 2024), and Qwen3 (1.7B) (Yang et al., 2025), with LoRA (Hu et al., 2022) (rank 32, α=64, dropout 0.05) in `bfloat16` with gradient checkpointing, sequence length 1024, one epoch, AdamW (β_1=0.9, β_2=0.999, weight decay 0.01, ϵ=10^{-8}), cosine schedule with 10% warmup, and gradient clipping at 1.0. The effective batch size is 16 on every backbone, reached with per-device batch 4 and gradient accumulation 4 on Sarvam-1 and Llama-3.2 and with per-device batch 2 and accumulation 8 on Qwen3. Each run fits on a single 48 GB GPU, and the full MorphDPO run uses $6{,}346$ steps on Sarvam-1 and Llama-3.2 and $6{,}347$ on Qwen3, at a peak of 25, 34 and 27 GB respectively. Baselines that do not use counterfactual data train on the $44{,}160$ reference pairs ($2{,}760$ steps), while MorphDPO trains on reference plus counterfactual pairs ($101{,}538$ examples), which accounts for its larger step count. The three backbones share this configuration and the frozen hyperparameters below without any per-backbone retuning, which leaves the backbone as the only intended difference between the three blocks of Table 1, with the reference-pair shuffle seed above as the one exception.

Frozen configuration. The configuration is selected by Optuna (Figure 6) on the validation split and frozen across all full runs and ablations, namely learning rate 4.2×10^{-5}, β=0.061, λ_{PO}=2.43, η=1.0, τ_g=0.1, shrinkage ρ=10, λ_{par}=λ_{SFT}=0, and a margin scale of zero, which sets δ_e=0 on every edge. Morphology weights are bounded to $[w_{\min}, w_{\max}]$=$[1, 3]$, with the per-feature increment above the stem weight capped at 2, and MorphDPO uses the normalized reward and marginalizes over up to 3 analyses per response. The counterfactual reliability is γ_c=0.8 for the four substitution types and γ_c=0.7 for the word-order transposition, with up to 3 corruptions per source response and no human-verified pair. The search space included the $\log Z$ correction as a binary choice and the selected trial has the correction *off*, and every full-scale MorphDPO number in this paper is obtained without that correction. The ablation of §3.3 isolates the effect separately and places it inside the replicate spread of those subset runs.

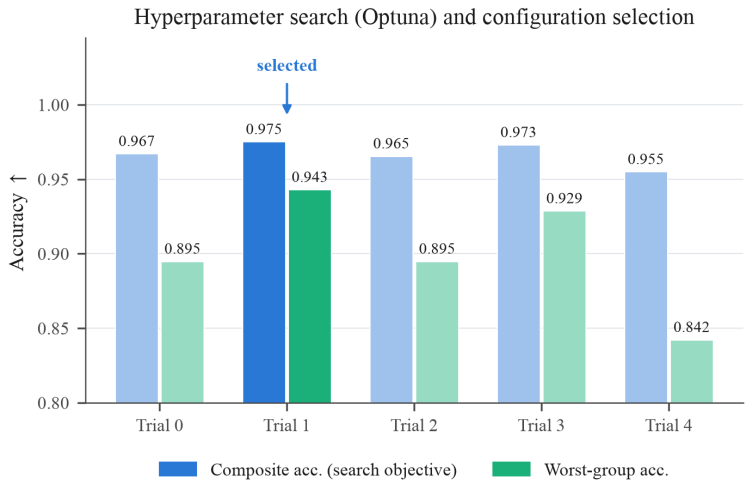


Figure 6: Optuna hyperparameter search. Composite accuracy (search objective) and worst-group accuracy for each trial, where the selected trial jointly maximizes both and is frozen for all full runs and ablations.

F Baseline Objectives

For completeness we state each baseline objective under our notation. Let $r_\theta(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_0(y|x)}$ be the DPO sequence log-ratio and $\bar{s}_\theta(x, y) = \frac{1}{|y|} \sum_t \log \pi_\theta(u_t \mid h_t)$ the mean token log-probability. All methods share the backbone, data budget, and evaluation of §3.1, together with the frozen learning rate and β of Appendix E, and the method-specific terms below are the only difference between runs.

DPO (Rafailov et al., 2023) minimizes a logistic loss on the sequence log-ratio between policy and frozen reference, the reference point from which all other objectives depart,

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\big(\beta[r_\theta(x, y^+) - r_\theta(x, y^-)]\big).$$

The log-ratio is summed over response tokens without length normalization, hence long responses accumulate more evidence than short ones. DPO is the strongest baseline on Sarvam-1, where composite accuracy of 0.9394 trails MorphDPO by 0.2 points, and the objective attains the highest reference-pair accuracy of any method on all three backbones. The tail behaves differently, since worst-group win-rate falls to 0.2766 on Llama-3.2 and 0.3617 on Qwen3, close to 11 points below MorphDPO on each.

SimPO (Meng et al., 2024) is reference-free and replaces the log-ratio with a length-normalized mean log-probability and adds a target margin γ,

$$\mathcal{L}_{\text{SimPO}} = -\log \sigma\big(\beta\, \bar{s}_\theta(x, y^+) - \beta\, \bar{s}_\theta(x, y^-) - \gamma\big).$$

Our runs set γ=0, hence the objective reduces to a logistic loss on the difference of mean log-probabilities. Removing π_0 drops the frozen copy of the backbone from memory yet forgoes the reference anchor, which we find correlates with weaker tail morphology, visible in a worst-group win-rate of 0.6809 on Sarvam-1 against 0.7660 for MorphDPO. The anchor matters less once the backbone is not Indic-centric, since SimPO overtakes DPO on composite accuracy on both general-purpose backbones, 0.7765 on Llama-3.2 and 0.8085 on Qwen3.

ORPO (Hong et al., 2024) is a monolithic, reference-free objective that adds an odds-ratio penalty to the supervised loss,

$$\mathcal{L}_{\text{ORPO}} = \mathcal{L}_{\text{SFT}}(x, y^+) - \lambda \log \sigma\Big(\log \tfrac{o_\theta(x,y^+)}{o_\theta(x,y^-)}\Big),$$

where $o_\theta(x, y) = p_\theta(y \mid x)/(1 - p_\theta(y \mid x))$ is the odds of generating y and p_θ is length-normalized. The frozen configuration carries λ_{SFT}=0 and λ_{PO}=2.43 into every run, hence the odds-ratio term

acts alone and the supervised anchor on y^+ drops out. ORPO ranks third on Sarvam-1 with composite accuracy 0.9290 and reaches the best honorificity accuracy of any method there, 0.9918, while number agreement stays at 0.8791 against 0.9372 for MorphDPO, the pattern of an objective that separates whole responses well and resolves individual features less reliably.

KTO (Ethayarajh et al., 2024) draws on prospect theory and does not require paired data. With reference point $z_0 = \mathrm{KL}(\pi_\theta \,\|\, \pi_0)$ and $r_\theta(x, y)$ as above, desirable (+) and undesirable (−) responses receive values $v_+(x, y) = \sigma\big(\beta(r_\theta(x, y) - z_0)\big)$ and $v_-(x, y) = \sigma\big(\beta(z_0 - r_\theta(x, y))\big)$, and the loss maximizes these gains with class weights λ_y,

$$\mathcal{L}_{\mathrm{KTO}} = \mathbb{E}\big[\lambda_y \,(1 - v_y(x, y))\big].$$

Our implementation converts each preference pair into two unary labels, the chosen response desirable and the rejected response undesirable, with equal class weights $\lambda_y{=}1$, length-normalized log-ratios, and z_0 estimated by the detached batch mean of the log-ratios clamped below at zero. Discarding the pairing costs the most on morphology, since minimal-pair accuracy falls to 0.8092 and worst-group win-rate to 0.4531 while reference-pair accuracy holds at 0.9889, and wrong case degrades furthest, to 0.5658 against 0.9865 for MorphDPO.

TDPO (Zeng et al., 2024) sharpens the training signal with a token-level sequential forward-KL regularizer $D_{\mathrm{SeqKL}}(x, y)$ between π_θ and π_0,

$$\mathcal{L}_{\mathrm{TDPO}} = -\log \sigma\big(\beta\,\Delta r - \beta\,[D_{\mathrm{SeqKL}}(x, y^-) - D_{\mathrm{SeqKL}}(x, y^+)]\big),$$

with $\Delta r = r_\theta(x, y^+) - r_\theta(x, y^-)$. Per-token divergence from the reference is controlled, yet the unit remains the model token. Our runs use the uniform-weight form of this objective, averaging the per-token log-ratio over the response and omitting the sequential-KL term, which isolates the effect of moving the unit from the sequence to the token. Averaging over tokens helps the general-purpose backbones, where TDPO is the strongest baseline on composite accuracy, 0.7789 on Llama-3.2 and 0.8103 on Qwen3, while the same objective ranks fifth on Sarvam-1 at 0.8951.

TIS-DPO (Liu et al., 2025) weights each token's log-ratio by an estimated importance $w_t \geq 0$, giving reward and loss

$$R_\theta(x, y) = \sum_t w_t \log \tfrac{\pi_\theta(u_t|h_t)}{\pi_0(u_t|h_t)}, \qquad \mathcal{L}_{\mathrm{TIS}} = -\log \sigma\big(\beta[R_\theta(x, y^+) - R_\theta(x, y^-)]\big).$$

Like MorphDPO, TIS-DPO reweights tokens, yet the weights are token-importance estimates rather than morphology-feature weights, and the reward is not renormalized into a valid log-ratio, the degeneracy quantified in §2.3. Importance in our runs follows reference surprisal, $w_t \propto \exp(-\log \pi_0(u_t \mid h_t)/T)$ with temperature $T{=}1$ and weights normalized to sum to one over the response, which concentrates credit on tokens the reference finds unlikely. Rare morphology and rare content share that property, hence the estimator spreads weight across both, and TIS-DPO ends up weakest on case, gender, and number agreement on the two general-purpose backbones, between 0.31 and 0.46.

MorphTok-DPO (Brahma et al., 2025) keeps the DPO objective and replaces only the subword tokenizer with a morphologically grounded one T_{morph}, i.e. $\mathcal{L}_{\mathrm{DPO}}$ evaluated on $u = T_{\mathrm{morph}}(y)$. This tests whether morphology-aware *tokenization* alone, without changing the objective, suffices. The analyzer inserts morpheme boundaries ahead of tokenization while the backbone tokenizer and embeddings stay in place, hence the change is confined to segmentation. Segmentation alone helps where morpheme boundaries align with existing subwords, visible in word-order accuracy of 0.9457 on Llama-3.2 and 0.9446 on Qwen3, and falls short on the categories that need feature-level credit, with number agreement at 0.3767 and 0.4767 and the lowest worst-group win-rate of any objective on either backbone. Our results indicate that tokenization alone does not suffice.

G Full Results

This appendix reports the per-method numbers behind the main-paper figures on all three backbones. The Llama-3.2 (3B) (Dubey et al., 2024) and Qwen3 (1.7B) (Yang et al., 2025) runs use the configuration of Appendix E with the identical frozen hyperparameters, morphology weights, and training

data as the Sarvam-1 runs. The corrupted half of the held-out test set is byte-identical across backbones, while the reference half redraws its shuffled negatives per backbone (Appendix E), hence the reference-pair column compares backbones on different negatives and the morphology columns do not. ORPO and KTO were run on Sarvam-1 only, and the other five baselines together with MorphDPO appear on every backbone. Each table stacks the backbones for one experiment. Training dynamics and the hyperparameter search appear in Appendix H, and the radar view of the per-language results appears in Appendix H.1.

Main Comparison. Table 1 of the main paper holds the primary comparison. Absolute scores on the two general-purpose backbones fall well below Sarvam-1, whose pretraining is Indic-centric, with worst-group win-rates of 0.38 and 0.47 against 0.77. MorphDPO nonetheless leads on composite accuracy, minimal-pair accuracy, worst-group win-rate, group gap, and overall win-rate on both backbones, DPO keeps a marginal lead on reference-pair accuracy, and TIS-DPO and MorphTok-DPO remain the weakest on the tail. The ranking among the baselines does shift, since DPO is second on Sarvam-1 and fourth on both general-purpose backbones, behind TDPO and SimPO. The morphology tax is larger on these backbones than on Sarvam-1, since the baselines lose between 2 and 20 points of minimal-pair accuracy relative to MorphDPO, and Figure 8 places MorphDPO at the top right of each panel. Reference-pair accuracy saturates between 0.95 and 0.97 on both general-purpose backbones, hence the composite margin of 2.8 to 3.3 points over the strongest baseline rests almost entirely on minimal pairs and on the tail. MorphTok-DPO makes that split explicit, pairing reference-pair accuracy near 0.97 with the lowest worst-group win-rates on either backbone, 0.15 on Llama-3.2 and 0.26 on Qwen3. The bottom of the ranking swaps as well, since MorphTok-DPO is last on Sarvam-1 at 0.7483 while TIS-DPO is last on both general-purpose backbones at 0.7050 and 0.7245. Group gaps stay wide throughout, where even MorphDPO reaches only 0.6170 on Llama-3.2 and 0.5304 on Qwen3 against 0.2340 on Sarvam-1.

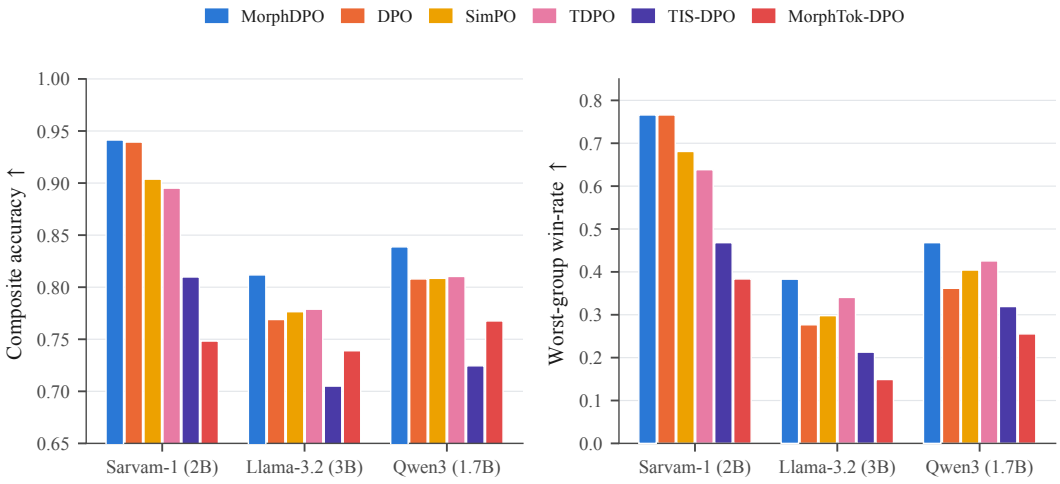


Figure 7: Composite accuracy (left) and worst-group win-rate (right) of the six shared objectives on each backbone, with MorphDPO first in every group.

Corruption Types. Table 3 lists the per-category accuracies plotted in Figure 3. Group sizes shape what each column can separate. Word order carries 4,293 of the 6,458 corrupted pairs, and every objective on Sarvam-1 except MorphTok-DPO scores above 0.90 there, hence the largest category contributes the least separation between methods. The word-order category also differs in kind from the other four, since the edit transposes two adjacent words rather than rewriting a slot inside one, which makes the category a test of surface order rather than of feature-level well-formedness. Number agreement carries only 430 pairs and holds the lowest MorphDPO accuracy of the five categories on Sarvam-1 and Llama-3.2. Honorificity rests on 122 pairs, where single-point differences stay within sampling noise. MorphTok-DPO is the only entry below 0.70 on word order, where morphology-aware pre-tokenization disturbs the token boundaries that carry the ordering signal.

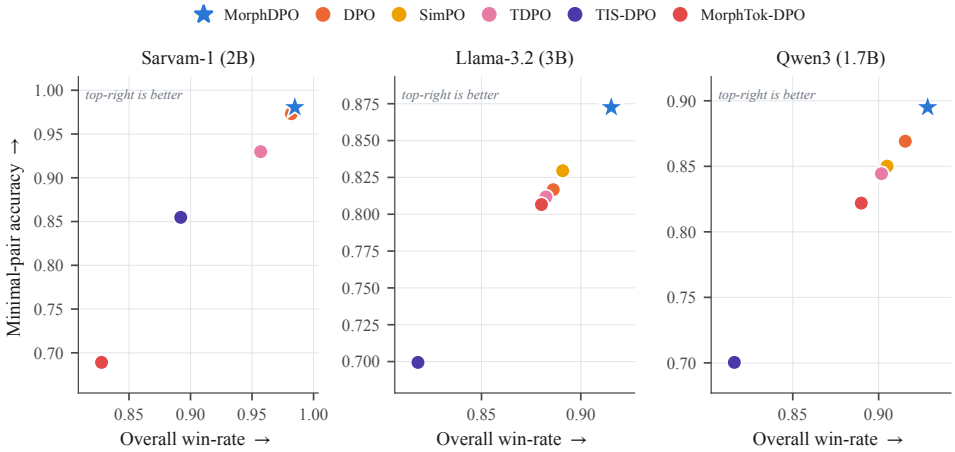


Figure 8: Overall win-rate against minimal-pair accuracy on each backbone, with MorphDPO (star) at the top right of every panel.

Table 3: Per-corruption-type accuracy on the held-out morphology test set for every backbone, with the best per column and backbone in **bold** and group sizes n in the header.

Backbone	Method	Wrong Case (n=889)	Gender Agr. (n=724)	Number Agr. (n=430)	Honorificity (n=122)	Word Order (n=4,293)
Sarvam-1 (2B)	**Ours**	**0.9865**	**0.9779**	**0.9372**	0.9836	0.9839
	DPO	0.9786	0.9530	0.8651	0.9672	**0.9865**
	ORPO	0.9595	0.9392	0.8791	**0.9918**	0.9856
	SimPO	0.8628	0.8163	0.7814	0.9098	0.9786
	TDPO	0.8853	0.8163	0.7674	0.8934	0.9755
	KTO	0.5658	0.6478	0.5674	0.7377	0.9131
	TIS-DPO	0.7987	0.6961	0.7000	0.8525	0.9087
	MorphTok-DPO	0.6625	0.6975	0.6581	0.7705	0.6939
Llama-3.2 (3B)	**Ours**	**0.7435**	**0.6657**	**0.6628**	**0.8934**	**0.9546**
	DPO	0.6828	0.5994	0.5163	0.8361	0.9106
	SimPO	0.6670	0.5470	0.5674	0.7951	0.9380
	TDPO	0.6434	0.4682	0.5465	0.6803	0.9348
	TIS-DPO	0.3881	0.3149	0.3465	0.5574	0.8682
	MorphTok-DPO	0.5579	0.5483	0.3767	0.7705	0.9457
Qwen3 (1.7B)	**Ours**	**0.8031**	**0.6989**	**0.7326**	**0.9180**	**0.9627**
	DPO	0.7762	0.6575	0.5349	0.8689	0.9576
	SimPO	0.7087	0.5691	0.5698	0.8115	0.9560
	TDPO	0.7109	0.5345	0.5767	0.7541	0.9536
	TIS-DPO	0.4567	0.3923	0.3907	0.5820	0.8372
	MorphTok-DPO	0.5849	0.5925	0.4767	0.8115	0.9446

Component Ablations. Table 4 repeats the controlled subset ablations of §3.3 on the two general-purpose backbones. The three effects reported on Sarvam-1 do not replicate here. Posterior marginalization and the exact $\log Z$ correction land within a point of their alternatives on composite accuracy on both backbones, and the unnormalized reward scores higher on composite accuracy than the normalized reward on both, by 2.0 points on Llama-3.2 and 2.1 points on Qwen3, while the worst-group effect of normalization is negative on Llama-3.2 and positive on Qwen3. Each block is again an independent set of runs, and the duplicate configuration of §3.3, namely the no-correction arm of the $\log Z$ block against the normalized arm of the reward block, separates by 0.9 points of composite accuracy and 4.3 points of worst-group win-rate on Llama-3.2 and by 0.05 and 0.0 points on Qwen3, which bounds the replicate spread on each backbone. These subset-scale runs also hold the Sarvam-1-tuned configuration fixed. The full-objective advantage of Table 1 is therefore the cross-backbone finding we stand behind, and the attribution to individual components is established on Sarvam-1 only, for reward normalization alone.

Table 4: Component ablations on the held-out morphology test set for the two general-purpose backbones, one factor varied per block over independent subset runs. Best per block in **bold**.

Backbone	Factor	Variant	Composite Acc. ↑	Reference-Pair Acc. ↑	Morph Min-Pair Acc. ↑	Worst-Group Win-Rate ↑	Group Gap ↓
Llama-3.2 (3B)	Analysis posterior	Hard 1-best	**0.7526**	**0.9406**	0.8238	**0.2340**	**0.7641**
		Posterior (ours)	0.7482	0.9397	**0.8244**	0.2128	0.7853
	log Z normalization	None	**0.7564**	0.9395	0.8238	**0.2553**	**0.7428**
		Approximate	0.7528	0.9402	0.8249	0.2340	0.7645
		Exact (ours)	0.7530	**0.9404**	**0.8250**	0.2340	0.7645
	Reward normalization	Unnormalized	**0.7674**	**0.9553**	**0.8249**	**0.2766**	**0.7215**
		Normalized (ours)	0.7478	0.9409	0.8222	0.2128	0.7853
Qwen3 (1.7B)	Analysis posterior	Hard 1-best	**0.7599**	0.9062	**0.7702**	**0.4468**	**0.5480**
		Posterior (ours)	0.7551	**0.9076**	0.7673	0.4255	0.5693
	log Z normalization	None	**0.7567**	**0.9067**	**0.7724**	**0.4255**	**0.5693**
		Approximate	0.7564	**0.9067**	0.7716	**0.4255**	**0.5693**
		Exact (ours)	0.7512	0.9062	0.7696	0.4043	0.5905
	Reward normalization	Unnormalized	**0.7771**	**0.9514**	**0.8212**	0.3404	0.6548
		Normalized (ours)	0.7562	0.9067	0.7710	**0.4255**	**0.5693**

H Training Dynamics and Per-Language Results

Figure 11 reports four headline metrics of Table 1 side by side on Sarvam-1, and Figure 12 traces the training dynamics of MorphDPO, where the loss decomposes into its active coefficient-weighted terms and the implicit reward separates chosen from rejected responses while training accuracy approaches one. Figure 9 plots the component ablations of Table 2, one panel per design choice.

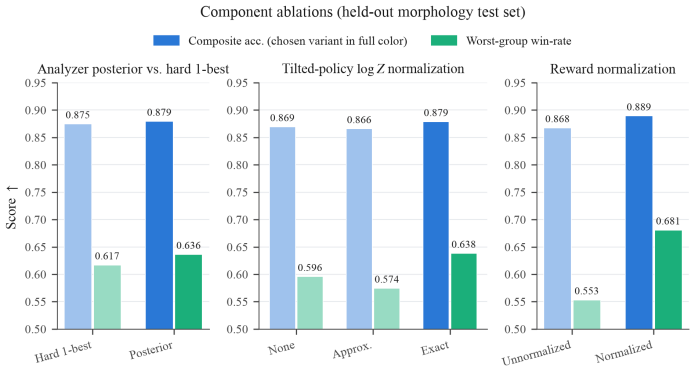


Figure 9: Component ablations on Sarvam-1. Each panel varies one design choice over an independent pair or triple of subset runs.

H.1 Per-Language Behavior

Figure 10 plots the Sarvam-1 column of Figure 4 as a radar profile over the eight languages. MorphDPO encloses the largest area, and its profile is close to a regular octagon, which is the visual signature of uniform quality across languages. Its shortest radius is Marathi at 0.96, and no axis falls below that. A method that trades one language against another produces a lobed profile instead, and SimPO does exactly that, pulling inward to 0.91 on Bengali and Telugu and 0.92 on Marathi while holding 0.99 on Hindi and Gujarati. DPO falls between the two, matching our profile on most axes and dipping on Telugu. The radar and the heatmap of Figure 4 carry the same numbers, and the radar makes the uniformity comparison immediate while the heatmap keeps every per-method value legible.

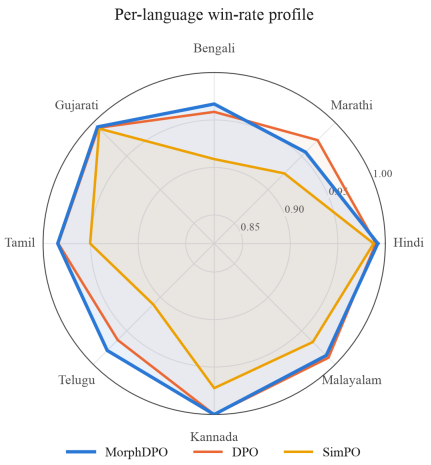


Figure 10: Per-language radar profile against two baselines.

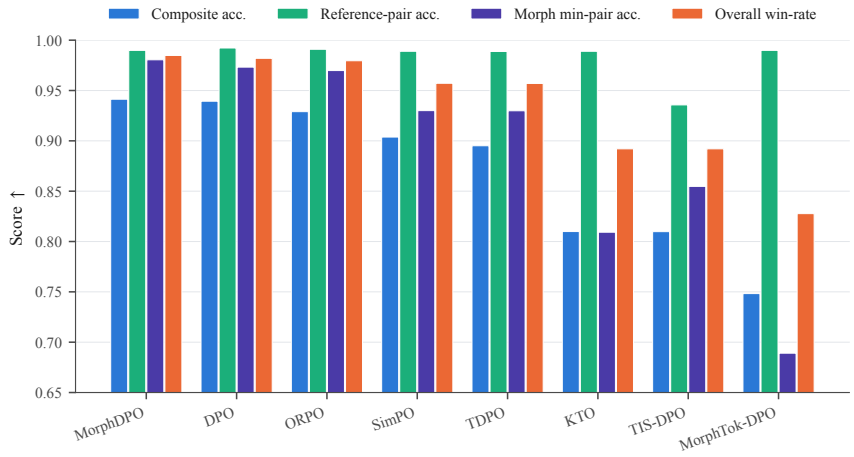


Figure 11: Primary metrics across preference-optimization objectives (composite, reference-pair, and morphology minimal-pair accuracy, and overall win-rate).

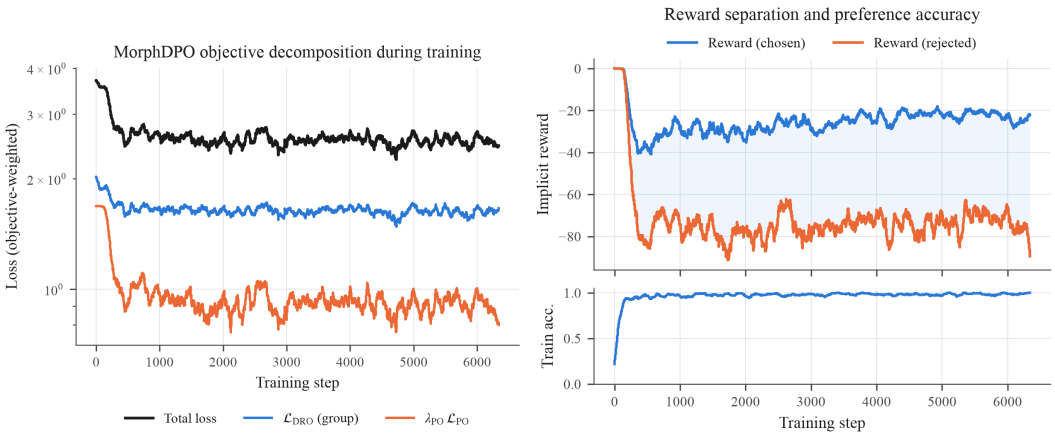


Figure 12: Training dynamics of MorphDPO. Left: the objective decomposes into its non-zero, coefficient-weighted terms, which sum to the total loss. Right: the implicit reward separates chosen from rejected responses as training preference accuracy approaches one.

I Analyses

Every example and count below is taken from the per-example prediction dumps generated during the experiments. Margins appear in scaled form, where the raw margin of a pair is divided by the margin scale of its own method and backbone over all 11,978 test pairs. A scaled margin $\sigma > 0$ therefore marks a correct ranking, and $|\sigma|$ states how decisive that ranking is in units of the objective's own typical margin. Raw margins carry each objective's own reward scale and do not compare across rows.

I.1 Qualitative Analysis

Table 5 collects corrupted minimal pairs where the morphology-aware objective changes the verdict. Each row is the arg-max of a stated predicate over the dumps, namely one corruption family together with a backbone on which MorphDPO ranks the preferred response first while DPO does not, and the two responses in a row are identical outside the marked span.

Table 5: Corrupted minimal pairs where MorphDPO ranks the preferred response first and DPO does not. Scaled margins σ are signed, where positive marks a correct ranking.

Language	Phenomenon	Backbone	Preferred → corrupted	σ DPO	σ Ours
Marathi	Gender agreement	Llama-3.2 (3B)	तीच → तोच	-0.009	$\mathbf{+1.224}$
Marathi	Number agreement	Llama-3.2 (3B)	आहेत → आहेतत	-0.032	$\mathbf{+0.992}$
Tamil	Case marking	Llama-3.2 (3B)	இனிக்கும் → இனிஇல்ம்	-0.013	$\mathbf{+0.960}$
Hindi	Honorificity	Llama-3.2 (3B)	प्रतियोग्यता → प्रतियिएग्यता	-0.014	$\mathbf{+0.885}$
Telugu	Number agreement	Qwen3 (1.7B)	నిలుపుకోవడం → నిపుకోవడం	-0.001	$\mathbf{+0.764}$
Bengali	Number agreement	Sarvam-1 (2B)	অপেক্ষকটির → অপেক্ষকগুলির	-0.017	$\mathbf{+0.263}$

The Marathi gender contrast turns on a single vowel sign, and MorphDPO separates the pair by more than a full margin scale while DPO stays undecided. The Bengali row is the one edit that stays inside the licensed inventory, since the singular classifier turns into its plural counterpart rather than into a non-word, and the margin there is the smallest of the six. Every other row edits a slot inside the word and produces a form outside the paradigm. Across all six rows the DPO scaled margins fall between -0.032 and -0.001, hence DPO is not confidently wrong on these pairs. The sequence-level reward simply carries too little signal at the deciding position, which is the behavior §2.3 predicts once a single morpheme is diluted across hundreds of tokens.

I.2 Error Analysis

Table 6 counts the decisions on which MorphDPO and each baseline actually disagree. Agreement carries no information about the objective, whereas b and c isolate the pairs that one objective ranks correctly and the other does not.

Table 6: Paired discordance against MorphDPO on the held-out set, where b counts pairs only MorphDPO ranks correctly and c counts pairs only the competitor does. Exact McNemar test.

Backbone	Competitor	b (ours only)	c (theirs only)	Δ accuracy
Sarvam-1 (2B)	DPO	117	83	$+0.0028$
	SimPO	384	53	$+0.0276$
	TDPO	388	55	$+0.0278$
	TIS-DPO	1,200	89	$+0.0928$
	MorphTok-DPO	1,963	81	$+0.1571$
Llama-3.2 (3B)	DPO	423	73	$+0.0292$
	SimPO	354	61	$+0.0245$
	TDPO	449	55	$+0.0329$
	TIS-DPO	1,230	64	$+0.0973$
	MorphTok-DPO	498	77	$+0.0351$
Qwen3 (1.7B)	DPO	235	79	$+0.0130$
	SimPO	338	55	$+0.0236$
	TDPO	380	58	$+0.0269$
	TIS-DPO	1,413	65	$+0.1125$
	MorphTok-DPO	543	80	$+0.0387$

Every comparison favors the morphology-aware objective at $p < 0.001$, apart from DPO on Sarvam-1 at $p = 0.019$, and the ratio b/c stays above 1.4 throughout. The two widest gaps name the two failure modes of the baselines, since token-importance weighting misses $1{,}200$ to $1{,}413$ pairs that MorphDPO ranks correctly, and morphology-aware tokenization alone misses $1{,}963$ on the Indic-specialized backbone. Table 7 turns to the pairs MorphDPO itself misranks.

Table 7: Pairs that MorphDPO misranks while DPO ranks correctly, covering four of the five corruption families.

Language	Phenomenon	Backbone	Preferred → corrupted	σ DPO	σ Ours
Hindi	Honorificity	Llama-3.2 (3B)	आप → तू	$+0.001$	-0.140
Marathi	Gender agreement	Sarvam-1 (2B)	तीन → तोन	$+0.007$	-0.089
Telugu	Case marking	Llama-3.2 (3B)	లోదుదు → కుదుదు	$+0.007$	-0.634
Malayalam	Word order	Llama-3.2 (3B)	കരുതപ്പെടുന്നു moved	$+0.111$	-0.634

MorphDPO misranks $1{,}014$ of the $11{,}978$ pairs on Llama-3.2 and 182 on Sarvam-1, with a mean scaled margin on error of -0.011 and -0.035 and no confident error, meaning no error with $\sigma < -1$, on either backbone. The failures are therefore undecided rather than actively learned. A further

892 of the Llama-3.2 errors are exact reward ties, 7.4% of the test set, where the corruption falls beyond the truncation window and no objective can separate the pair, and excluding those ties moves accuracy from 0.915 to 0.989. The honorificity row is the most instructive of the four, since the corruption replaces an honorific pronoun with a familiar one, both of them licensed forms, hence the pair asks for register knowledge rather than well-formedness. The Marathi row carries the same vowel-sign contrast that MorphDPO resolves correctly on Llama-3.2 in Table 5, which places the residual errors inside the categories the objective otherwise handles.

J LICENSE AND ASSETS

We build on publicly released models, datasets, and software and use each in accordance with its license, and users should re-verify terms before use. The backbone Sarvam-1 (Sarvam AI, 2024) is used under the terms of its public model release, Llama-3.2 (3B) (Dubey et al., 2024) under the Llama 3.2 Community License, and Qwen3 (1.7B) (Yang et al., 2025) under Apache-2.0. Parameter-efficient fine-tuning uses LoRA/PEFT (Apache-2.0), and hyperparameter search uses Optuna (MIT). Morphological feature schemas are from UniMorph (Batsuren et al., 2022) (CC-BY-SA). The MorphPref-Indic preference and counterfactual data introduced here are derived from openly available Indic instruction resources (Khan et al., 2024), and we will release the construction code, morphology-transformation library, and evaluation splits under the Creative Commons Attribution 4.0 International license (CC BY 4.0), subject to the terms of the underlying sources.

K BROADER IMPACT

MorphDPO aims to make aligned language models more reliable in morphologically rich, often lower-resource languages, where fluent-looking yet grammatically incorrect output can mislead users, entrench a majority-dialect bias, or degrade access to information for speakers of underserved languages. By optimizing worst-group performance across language-script-morphology groups rather than average win-rate, the method is explicitly designed to reduce disparities across languages and scripts rather than amplify them. The same machinery, however, carries the usual risks of preference optimization. The notion of "correct" morphology encoded by the analyzer and the counterfactual transformations reflects particular standard varieties and annotation choices, and could marginalize non-standard yet legitimate dialectal or sociolinguistic forms if applied uncritically. Morphology-counterfactual supervision could also, in principle, be repurposed to make generation more persuasive or harder to detect. We mitigate these concerns by keeping morphology weights bounded and fixed, by requiring a validity predicate and human spot-checks on synthetic pairs, and by reporting stratified, worst-group results, which keeps degradation on any language or register visible rather than hidden in an average. We recommend that deployments document the analyzer, schema, and dialects used, and involve native speakers in validating the notion of acceptability being optimized.

L AI USE STATEMENT

Generative AI tools used in this work as general-purpose assistants under our own supervision. Their use stayed within two limits, the polishing of prose we had already drafted, covering grammar, phrasing, and formatting, and routine engineering such as plotting scripts and refactoring of the experiment harness. These tools were not used to create the idea for the research or to implement the core methodology. We reviewed every AI-assisted passage and every AI-assisted line of code, and each number in this paper comes from the code released, run by us.