GapPO: Gradient-Adaptive Pairwise Preference Optimization

Anonymous Authors[1]

Abstract

Aligning large language models to human preferences requires training on pairwise comparisons between candidate responses. Existing preference optimization methods assign equal gradient weight to every pair, regardless of whether the quality difference is large or negligibly small. We introduce GapPO (Gradient-Adaptive Pairwise Preference Optimization), a preference optimization method designed to directly improve pairwise ranking accuracy in large language models. Standard methods currently treat all pairs equally: a pair scoring 4.8 vs. 1.2 receives the same gradient weight as one scoring 3.2 vs. 2.9, diluting clear signal with annotator noise. GapPO corrects this by weighting each pair by the absolute quality-score gap $|\delta| = |\texttt{score}_{\text{chosen}} - \texttt{score}_{\text{rejected}}|$, so that gradient mass concentrates on the most discriminative comparisons. Since the model is shaped more by reliable comparisons, its implicit reward function better separates high-quality from low-quality responses at test time. Beyond improving pairwise accuracy (PWA), score-gap weighting improves Spearman rank correlation between model rewards and annotation scores, which is the calibration property required to scale from pairwise to listwise ranking. Evaluated on UltraFeedback binarized across Qwen2.5-0.5B, Gemma-2-2B, and Mistral-7B, GapPO consistently outperforms SimPO, CPO, IPO, and AlphaPO baselines.

1. Introduction

Ranking is a core problem in document retrieval (Manning et al., 2008; Liu, 2009) and language model alignment (Christiano et al., 2017; Ouyang et al., 2022). Given a query and a set of candidate responses, a model must assign implicit scores that correctly order them by quality. Large language models (LLMs) are increasingly being used as rankers (Sun et al., 2023; Hou et al., 2024; Qin et al., 2024), but standard preference optimization methods are designed for general alignment, not specifically targeting ranking accuracy. They optimize a binary preference signal (chosen vs. rejected) and treat every pair with the same gradient weight, regardless of how different the two responses actually are.

This uniformity is a poor fit for ranking. A pair where one response scores 4.8 and another scores 1.2 is an unambiguous signal where any reasonable judge would agree on the ordering. Conversely, a pair with scores 3.2 and 2.9 is near a tie, which suggests that the label may reflect annotator noise more than a genuine quality difference. Training a ranker on both pairs with equal weight dilutes the strong signal with the weak. The model spends as much gradient on borderline cases as on clear-cut ones.

GapPO corrects this by deriving a per-pair training weight directly from the quality-score gap $|\delta^{(i)}| = |s_w^{(i)} - s_l^{(i)}|$, where s_w and s_l are the human annotator scores for the chosen and rejected responses. Pairs with large $|\delta|$ receive proportionally more gradient while pairs with small $|\delta|$ are down-weighted but not discarded. This is structurally different from the alternative of hard filtering (discarding all pairs with $|\delta| < \tau$). The continuous weight in GapPO preserves moderate-$|\delta|$ pairs at reduced contribution, while a hard filter treats them identically to high-$|\delta|$ pairs after thresholding, discarding useful signal below the cutoff.

GapPO introduces score-gap weighting as a general mechanism that can be applied to any per-sample preference loss. We build on SimPO's reference-free, length-normalized reward (Meng et al., 2024) — $r(y|\mathbf{x}) = \frac{1}{|y|} \log \pi_\theta(y|\mathbf{x})$ — and multiply each pair's loss by $|\delta^{(i)}|$, shifting gradient mass toward the most reliable comparisons. Weights are left unnormalized, so batches rich in high-$|\delta|$ pairs produce larger update steps, allowing the training signal to be strongest when the evidence is clearest.

The implicit reward $r(y|\mathbf{x})$ learned during preference optimization is a ranking function: it assigns a score to every response, and those scores induce an ordering. Pairwise accuracy (PWA), which is the fraction of held-out pairs the model correctly ranks, directly measures whether that

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

ordering agrees with human preferences. While PWA is necessary, correctly ordering isolated pairs does not guarantee a globally consistent ranking across all responses to a prompt. We also measure Spearman rank correlation between model rewards and annotation scores, a continuous indicator of reward calibration and a direct bridge to listwise ranking, where all k responses must be jointly ordered. GapPO improves Spearman across all model families, suggesting score-gap weighting yields a more globally calibrated reward function, not just better pairwise comparisons.

Our main contributions are:

- **Identifying a fundamental gap in preference optimization**: existing methods treat all training pairs with equal gradient weight, ignoring the reliability signal embedded in annotation score differences. A pair with $|\delta| = 4.5$ is treated identically to one with $|\delta| = 0.3$, causing models to spend as much gradient on annotator noise as on unambiguous quality signal.
- **GapPO**: a novel preference optimization loss that directly addresses this gap by weighting each training pair by its annotation score gap $|\delta|$. GapPO concentrates gradient on the most reliable comparisons and weights are left unnormalized so that batch informativeness modulates the effective learning rate, allowing training to be strongest precisely when the evidence is clearest.
- **Consistent empirical gains across models and fine-tuning regimes**: GapPO outperforms SimPO, CPO, IPO, and AlphaPO on peak pairwise accuracy (PWA) and Spearman rank correlation across Qwen2.5-0.5B, Gemma-2-2B, and Mistral-7B under both full and MLP-only fine-tuning, demonstrating that score-gap weighting generalizes across model families, scales, and parameter-efficient trainings.

2. Preliminaries

2.1. Problem Definition

Let $\mathcal{D} = \{(\mathbf{x}^{(i)}, y_w^{(i)}, y_l^{(i)}, s_w^{(i)}, s_l^{(i)})\}_{i=1}^N$ be a preference dataset where $\mathbf{x}^{(i)}$ is a prompt, $y_w^{(i)}$ and $y_l^{(i)}$ are the chosen and rejected responses, and $s_w^{(i)}, s_l^{(i)} \in \mathbb{R}$ are their corresponding human annotation scores. Each trained policy π_θ induces an implicit reward function $r(y|\mathbf{x})$ that assigns a scalar score to every response. We define the score gap as:

$$\delta^{(i)} = \left| s_w^{(i)} - s_l^{(i)} \right|. \tag{1}$$

The goal is to learn π_θ such that the induced reward correctly ranks responses according to human preference. Formally, we aim to minimize the expected ranking error:

$$\mathcal{R}(\theta) = \mathbb{E}_{(\mathbf{x}, y_w, y_l)\sim\mathcal{D}}[\mathbf{1}[r(y_w|\mathbf{x}) \le r(y_l|\mathbf{x})]], \tag{2}$$

which is the complement of pairwise accuracy (PWA). Since this objective is non-differentiable, preference optimization methods minimize a surrogate loss. The key limitation of existing surrogates is that they weight every pair equally, making the optimization indifferent to $\delta^{(i)}$. Pairs where annotation scores are nearly identical carry as much gradient as pairs with large, unambiguous score gaps.

2.2. LLMs as Rankers

Using LLMs to rank candidate responses is a growing area of research (Sun et al., 2023; Ma et al., 2023). Most approaches either prompt a frozen LLM in a listwise or pointwise fashion, or fine-tune on binary preference labels. A key limitation of these approaches is that they treat the ranking problem as a classification task: the model learns to distinguish chosen from rejected responses without access to how large the quality difference between them actually is. GapPO addresses this by fine-tuning the implicit reward function directly on graded annotation scores, making ranking accuracy an explicit training objective rather than a byproduct of binary preference learning.

2.3. Reference-Free Preference Optimization

The dominant paradigm for aligning LLMs with human preferences is reinforcement learning from human feedback (RLHF) (Christiano et al., 2017), which trains a reward model on pairwise comparisons and then optimizes the policy via PPO (Schulman et al., 2017). DPO (Rafailov et al., 2023) simplified optimizes this policy on preference pairs without an explicit reward model, showing that the RLHF objective can be reparameterized as a supervised loss. Subsequent work has explored modifications to the DPO objective. IPO (Azar et al., 2024) replaces the logistic loss with a squared loss to avoid overfitting to hard labels. CPO (Xu et al., 2024) removes the reference model entirely while SimPO (Meng et al., 2024) introduces length normalization and a target reward margin to improve calibration. AlphaPO (Gupta et al., 2025) generalizes the reward shape with an α-parameterized transformation. However, all of these methods derive their training signal purely from the binary preference label (chosen vs. rejected), ignoring the magnitude of the underlying quality difference. GapPO introduces annotation score gaps as an explicit training signal, making gradient allocation a function of annotation confidence rather than label identity.

We build on reference-free preference objectives, which train the model without a frozen reference model at inference time. The key baselines are formalized below.

CPO (Xu et al., 2024) applies a hinge loss on raw log-

probabilities:

$$\mathcal{L}_{\text{CPO}}(\theta) = -\mathbb{E}\Big[\log\sigma\Big(\beta\big(\log\pi_\theta(y_w|\mathbf{x}) - \log\pi_\theta(y_l|\mathbf{x})\big) - \beta\Big)\Big]. \tag{3}$$

IPO (Azar et al., 2024) uses a squared loss to reduce overfitting to hard labels:

$$\mathcal{L}_{\text{IPO}}(\theta) = \mathbb{E}\left[\left(\log\frac{\pi_\theta(y_w|\mathbf{x})}{\pi_{\text{ref}}(y_w|\mathbf{x})} - \log\frac{\pi_\theta(y_l|\mathbf{x})}{\pi_{\text{ref}}(y_l|\mathbf{x})} - \frac{1}{2\beta}\right)^2\right]. \tag{4}$$

SimPO (Meng et al., 2024) uses a length-normalized log-probability reward $r(y|\mathbf{x}) = \frac{1}{|y|}\log\pi_\theta(y|\mathbf{x})$ with a target margin γ:

$$\mathcal{L}_{\text{SimPO}}(\theta) = -\mathbb{E}\big[\log\sigma\big(\beta\big(r(y_w|\mathbf{x}) - r(y_l|\mathbf{x})\big) - \gamma\big)\big]. \tag{5}$$

AlphaPO (Gupta et al., 2025) replaces the log reward with an α-parameterized transformation $r_\alpha(y|\mathbf{x}) = \frac{1}{|y|}\sum_t \frac{\pi_\theta(y_t|\mathbf{x})^\alpha - 1}{\alpha}$, modifying the reward shape while keeping pair weights uniform.

All four methods treat every preference pair with equal gradient weight, providing no mechanism to use the magnitude of the annotation score gap as a training signal.

2.4. Weighted and Filtered Preference Optimization

Several works have explored using data quality signals to improve preference learning. Zhou et al. (2023) weight pairs by annotator confidence scores derived from inter-annotator agreement; Liu et al. (2024) use rejection sampling to discard unreliable pairs before training. These approaches either require additional annotator metadata or resort to binary filtering. GapPO differs in using a differentiable continuous weight derived directly from annotation score gaps, preserving all training pairs while proportionally down-weighting noisy comparisons. It also differs in objective: prior work targets general alignment, whereas GapPO explicitly targets ranking accuracy as measured by PWA and Spearman correlation.

2.5. Parameter-Efficient Fine-Tuning

LoRA (Hu et al., 2022) and QLoRA (Dettmers et al., 2024) are the dominant parameter-efficient fine-tuning methods, decomposing weight updates into low-rank matrices. MLP-only fine-tuning, evaluated here, is a complementary structured approach motivated by the observation that feed-forward layers store factual knowledge (Geva et al., 2021) and may dominate the representational changes needed for preference adaptation. Unlike LoRA, MLP-only tuning modifies weights directly and requires no adapter merging at inference time.

3. Methodology: Gradient-Adaptive Pairwise Preference Optimization (GapPO)

GapPO is motivated by a simple observation: annotation score gaps are a direct proxy for how reliable a preference label is, yet no existing preference optimization method uses this signal to allocate gradient.

3.1. Score-Gap Weighting: Concentrating Gradient on Reliable Pairs

The central insight of GapPO is that human annotator score gaps are a direct proxy for pair reliability, and gradient allocation should reflect this. Let $s_w^{(i)}$ and $s_l^{(i)}$ be the quality scores for the chosen and rejected responses of the i-th training pair respectively. We define:

$$\delta^{(i)} = \left|s_w^{(i)} - s_l^{(i)}\right|. \tag{6}$$

We map $\delta^{(i)}$ to a non-negative weight via linear scaling:

$$w^{(i)} = \delta^{(i)}. \tag{7}$$

The GapPO objective is then:

$$\mathcal{L}_{\text{GapPO}}(\theta) = -\frac{1}{B}\sum_{i=1}^{B} w^{(i)}\log\sigma\Big(\beta\big(r(y_w^{(i)}|\mathbf{x}^{(i)}) - r(y_l^{(i)}|\mathbf{x}^{(i)})\big) - \gamma\Big), \tag{8}$$

where $r(y|\mathbf{x}) = \frac{1}{|y|}\log\pi_\theta(y|\mathbf{x})$ is a reference-free, length-normalized log reward. When all $\delta^{(i)}$ are equal, $w^{(i)}$ is constant and GapPO reduces to the baseline SimPO function. The weight $w^{(i)}$ appears *outside* the sigmoid, acting as a per-sample loss multiplier that rescales the gradient contribution of each pair without changing the functional form of the pairwise comparison.

Figure 1 illustrates the effect empirically. Baseline methods (CPO, IPO, SimPO, AlphaPO) concentrate gradient mass at low-to-moderate $|\delta|$ simply because that is where most UltraFeedback pairs lie. GapPO actively redistributes mass toward the high-$|\delta|$ tail, where annotation labels are most reliable.

Linear scaling is the natural choice among monotone weighting functions. It is the unique weight proportional to annotation confidence. A pair with $|\delta| = 4$ contributes exactly twice as much gradient as one with $|\delta| = 2$, directly reflecting how much more reliable the annotation is. Superlinear alternatives such as quadratic ($w = |\delta|^2$) or exponential ($w = e^{|\delta|} - 1$) concentrate mass so aggressively on the extreme high-$|\delta|$ tail that moderate-$|\delta|$ pairs contribute almost nothing, which defeats the stated purpose of soft weighting over hard filtering. Figure 2 shows this effect on the UltraFeedback $|\delta|$ distribution directly.

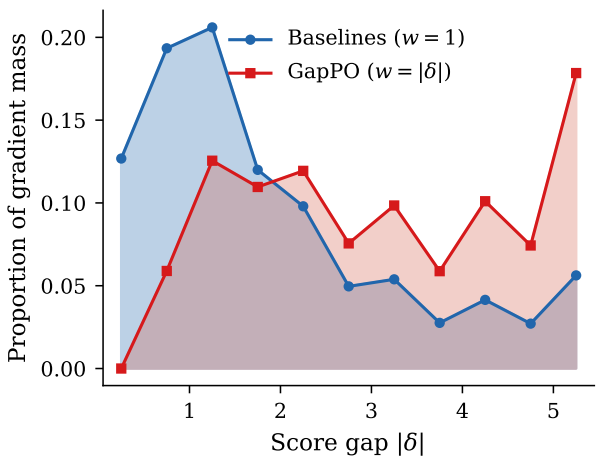

Figure 1. Fraction of total gradient mass per $|\delta|$ bin on UltraFeedback training pairs. Baselines apply uniform per-pair weights ($w = 1$), so their distribution mirrors the raw dataset histogram. GapPO ($w = |\delta|$) actively shifts mass toward high-confidence pairs.

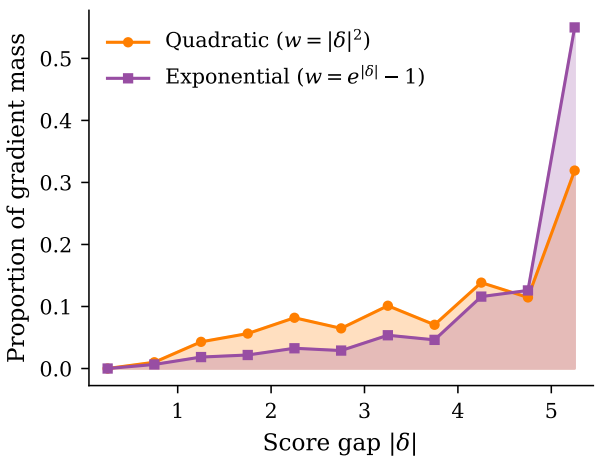

Figure 2. Gradient mass allocation under superlinear weighting functions. Quadratic ($w = |\delta|^2$) and exponential ($w = e^{|\delta|} - 1$) concentrate mass so heavily at the extreme high-$|\delta|$ tail that moderate-$|\delta|$ pairs contribute negligible gradient.

3.2. Unnormalized Weights as a Batch-Adaptive Learning Rate

We use the raw weights $w^{(i)} = \delta^{(i)}$ directly without batch normalization. This allows batches containing many high-$|\delta|$ pairs to produce proportionally larger gradient steps, creating a form of adaptive learning rate that provides batch informativeness. This is preferable to normalizing within each batch, which would reduce the differential and make the effective update magnitude insensitive to how informative the batch actually is.

3.3. Continuous Soft Weighting vs. Hard Pair Filtering

A natural alternative to GapPO's continuous weighting is hard filtering: discard all pairs with $\delta^{(i)} < \tau$ and train on the remainder with uniform weights. This is equivalent to a step-function weight $w^{(i)} = \mathbf{1}[\delta^{(i)} \geq \tau]$.

Hard filtering has two disadvantages compared to GapPO:

1. It applies a binary threshold, so all retained pairs receive equal gradient weight regardless of how large their score gap is. A pair with $\delta = 2.1$ is treated identically to one with $\delta = 4.8$. GapPO continues to differentiate within the retained set.

2. The threshold is a discrete hyperparameter that must be tuned, whereas GapPO's continuous weight adapts naturally to the full distribution of $|\delta|$ in the training data without requiring a cutoff decision.

We compare GapPO (continuous weighting) against a hard-filtered baseline (`min_delta`$= \tau$, uniform weights on retained pairs) directly in our experiments, using the same τ to ensure a fair comparison of filtering vs. continuous weighting strategies.

3.4. Role of the Target Reward Margin and Its Interaction with Delta Weighting

The margin γ in Equation 8 controls the minimum gap the model must maintain between chosen and rejected rewards before the loss saturates. In GapPO, γ interacts with delta weighting: pairs where the model already clears the margin contribute little loss regardless of δ, so the effective training signal is concentrated on pairs that are simultaneously (a) hard for the model to rank correctly and (b) unambiguous in annotation. We find $\gamma = 0.5$ provides a good balance, requiring a non-trivial margin while avoiding over-penalizing near-saturation pairs.

3.5. MLP-Only Fine-Tuning as a Parameter-Efficient Ranking Regime

We adopt a parameter-efficient strategy that freezes all attention sublayers, layer-norm parameters, and embedding matrices, training only the MLP (feed-forward network) sublayers. For the models in this work, this amounts to approximately 63.5% of total parameters. Unlike LoRA (Hu et al., 2022), MLP-only tuning modifies weights directly and requires no adapter merging at inference time. It is motivated by the observation that feed-forward layers store factual knowledge (Geva et al., 2021) and dominate representational change needed for preference adaptation. We contrast MLP-only tuning directly against full model fine-tuning across all methods, both as a more compute-efficient training regime and to assess whether the ranking

gains from delta weighting persist even when the majority of the model's parameters are held fixed.

3.6. Training Stability and Collapse Mitigation

Length-normalized objectives are susceptible to two failure modes that must be addressed at the methodology level. The first is a sharp crash triggered by fully-truncated sequences: when a prompt is long enough to consume the entire 1024-token budget, the response has zero valid tokens after tokenization, causing length normalization to divide by zero and produce NaN losses that corrupt gradients. We address this by clamping the response length denominator to a minimum of 1, so fully-truncated responses receive a reward of 0 rather than propagating NaN.

The second is a gradual collapse driven by a length bias in the training data: chosen responses tend to be longer on average than rejected responses. Length normalization removes the direct per-token advantage, but does not eliminate the subtler distributional correlation between response length and preference label. Over continued training, the model can exploit this correlation by assigning higher length-normalized reward to responses with the stylistic patterns of chosen outputs rather than learning genuine quality distinctions. As the model's output distribution narrows toward these patterns, per-token log-probabilities inflate and the reward loses discriminative power.

GapPO's delta weighting mitigates the second failure mode by design: high-$|\delta|$ pairs carry genuine quality differences that provide a corrective signal early in training, extending the window before the length-correlation collapse dominates. Additional mitigation strategies applied across all methods include: (1) clamping the length denominator to $\min = 1$; (2) 1-epoch training to stop before collapse; (3) cosine LR schedule with 10% warmup; (4) conservative learning rates; and (5) `load_best_model_at_end=True` to save the peak-PWA checkpoint rather than the final one.

4. Experiments

4.1. Training Details

All experiments use the **UltraFeedback binarized** dataset (Cui et al., 2023), which provides explicit scalar quality scores (`score_chosen` and `score_rejected`) for each preference pair, making δ directly available without additional preprocessing. Scores are continuous averages over multiple annotation criteria, yielding $|\delta| \in [0, 5.5]$.

We evaluate on four pretrained models spanning a range of scales, architectures, and training lineages: Qwen2.5-0.5B, Gemma-2-2B, and Mistral-7B. The original SimPO (Meng et al., 2024) and AlphaPO (Gupta et al., 2025) evaluations both use Llama-3-8B, Mistral-7B, and Gemma-2-9B; we include Mistral-7B for direct comparison, substitute Gemma-2-2B as a smaller representative of the Gemma-2 family, and add Qwen2.5-0.5B as an SLM to test whether GapPO's gradient concentration on high-$|\delta|$ pairs yields greater benefit at small scale. Each model is evaluated under both MLP-only and full fine-tuning regimes.

We compare GapPO against four baselines: **CPO** (Xu et al., 2024) (reference-free hinge loss), **IPO** (Azar et al., 2024) (squared loss), **SimPO** (Meng et al., 2024) (length-normalized, uniform pair weights), and **AlphaPO** (Gupta et al., 2025) (α-shaped reward, uniform pair weights, $\alpha = 0.5$). We also include a hard-filter baseline that discards pairs with $|\delta| < 1$ to contrast continuous weighting against hard pair selection.

All methods are implemented within the `CPOTrainer` framework (Xu et al., 2024) so that PWA is computed identically across methods. Learning rates are scaled by model size: 1×10^{-5} for Qwen2.5-0.5B, 8×10^{-6} for Gemma-2-2B, and 5×10^{-6} for Mistral-7B, with a cosine schedule and 10% linear warmup. We train for 1 epoch with batch size 2 per device and 8 gradient accumulation steps (effective batch size 16), maximum sequence length 1024 tokens, $\beta = 1.0$, $\gamma = 0.5$, and random seed 42.

4.2. Evaluation Metrics: Pairwise Accuracy and the Bridge to Listwise Ranking

We evaluate all methods on two complementary metrics. The primary metric is **pairwise accuracy (PWA)**:

$$\text{PWA} = \frac{1}{|\mathcal{D}_{\text{eval}}|} \sum_{(x, y_w, y_l) \in \mathcal{D}_{\text{eval}}} \mathbf{1}[r(y_w|x) > r(y_l|x)], \quad (9)$$

where $r(y|x) = \frac{1}{|y|} \log \pi_\theta(y|x) = \frac{1}{|y|} \sum_t \log \pi_\theta(y_t|x, y_{<t})$ is the token-averaged log-probability of a fixed response y under the trained policy. Critically, no generation occurs at evaluation time: both y_w and y_l from the held-out set are forward-passed through the model to obtain their scalar rewards, and the model is compared on how well it scores these existing responses rather than on what it would generate. PWA measures directly whether the model assigns a higher implicit reward to the human-preferred response in each held-out pair. The evaluation set is filtered to pairs with $|\delta| \geq 2$, restricting measurement to high-confidence annotations. Unlike open-ended generation benchmarks, PWA is deterministic and directly tied to the training objective. We report **peak PWA**, which is the maximum PWA achieved at any checkpoint during training instead of the final-step value, since methods that overfit or collapse late in training would be penalized for instability unrelated to their core optimization quality.

We additionally report **Spearman rank correlation** ρ between the model's implicit reward and annotation scores

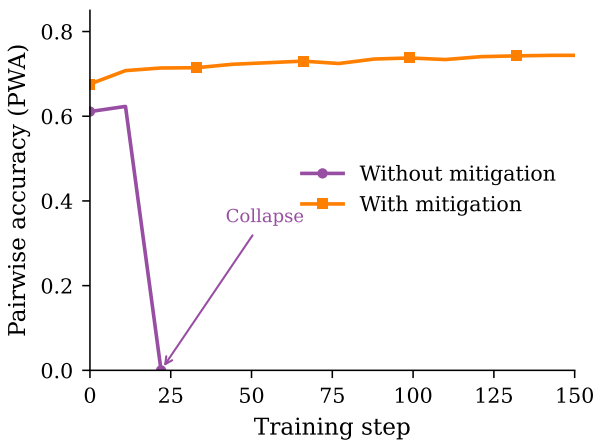

Figure 3. PWA over training steps with and without collapse mitigation on Qwen2.5-0.5B-Instruct.

across the full evaluation set. Let $\{(r_i, s_i)\}_{i=1}^{M}$ denote the implicit rewards and annotation scores for all M responses (chosen and rejected pooled) in the eval set. Let $\mathrm{rk}(r_i)$ and $\mathrm{rk}(s_i)$ be their rank positions. Spearman ρ is:

$$\rho = 1 - \frac{6\sum_{i=1}^{M} d_i^2}{M(M^2-1)}, \quad d_i = \mathrm{rk}(r_i) - \mathrm{rk}(s_i). \tag{10}$$

While PWA measures whether individual pairs are correctly ordered, Spearman measures whether the reward is globally calibrated as a continuous scoring function — the property required to scale from pairwise to listwise ranking. A model with high Spearman assigns scores that monotonically track annotation quality across the full quality spectrum, not just their direction, which is the prerequisite for strong listwise NDCG performance. Figure 5 shows Spearman trajectories over training.

4.3. Training Collapse Validation

Section 3.6 describes two failure modes of length-normalized objectives: NaN crashes from zero-length responses and gradual PWA degradation from length-label correlation. Figure 3 validates the second failure mode empirically. Without collapse mitigation, PWA rises briefly in the first few hundred steps before crashing to near zero as the model degenerates, which confirms that early stopping and best-checkpoint selection are necessary rather than optional. With mitigation strategies applied (length denominator clamping, 1-epoch training, cosine LR schedule, and `load_best_model_at_end`), PWA rises steadily and stabilizes, demonstrating that the failure mode is reproducible and addressable.

4.4. Results

Table 1 reports Spearman Correlation and peak PWA for all methods, models, and fine-tuning regimes. Table 2 reports Spearman correlation and PWA for Qwen2.5-0.5B-Instruct. Figures 4 and 5 show training curves.

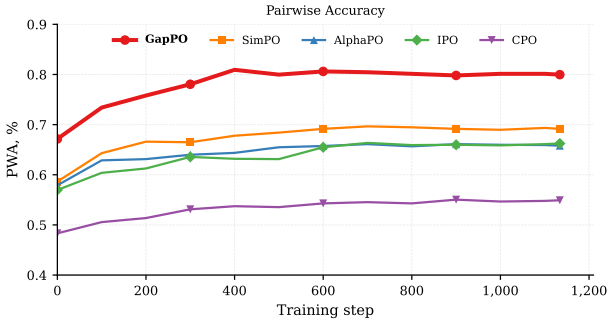

Figure 4. Pairwise accuracy (PWA) over training steps on Qwen2.5-0.5B-Instruct ($\beta = 1.0$, $\gamma = 0.5$, 1 epoch, cosine LR). All methods are evaluated on the held-out UltraFeedback split every 100 steps.

Score-gap weighting yields large, consistent PWA gains. GapPO improves peak PWA over the strongest baseline (SimPO) by $+11.28\%$ on Qwen2.5-0.5B MLP-only, $+11.39\%$ on Qwen2.5-0.5B full, $+4.46\%$ on Gemma-2-2B MLP-only, $+4.05\%$ on Gemma-2-2B full, $+0.64\%$ on Mistral-7B MLP-only, and $+0.82\%$ on Mistral-7B full. Gains are largest for small language models (Qwen2.5-0.5B), which is consistent with GapPO's gradient concentration being most consequential when the total model capacity is limited.

Score-gap weighting produces substantially higher Spearman correlation. GapPO improves Spearman over SimPO by $+15.99\%$ on Qwen2.5-0.5B MLP-only, $+15.44\%$ on Qwen2.5-0.5B full, $+6.64\%$ on Gemma-2-2B MLP-only, $+4.09\%$ on Gemma-2-2B full, $+0.45\%$ on Mistral-7B MLP-only, and $+1.23\%$ on Mistral-7B full. Higher Spearman indicates the reward is more globally calibrated across the full quality spectrum, which directly validates the core GapPO claim that weighting by $|\delta|$ shapes the reward to reflect annotation confidence rather than just binary preference direction. This is the calibration property required to scale to listwise NDCG evaluation.

MLP-only fine-tuning retains most of the gain at lower compute cost. Across all models, MLP-only tuning achieves PWA within 0.5–1.0% of full fine-tuning while training only $\sim$36.5% of parameters. On Gemma-2-2B, GapPO MLP-only reaches 85.26% vs. 86.22% for full fine-tuning, which is a 1.0% gap that is small relative to the 4.46% gain over SimPO. This confirms that the ranking gains persist under parameter-efficient training.

Table 1. Spearman Corr (%) and Peak pairwise accuracy (PWA, %) on UltraFeedback binarized eval set. All results use $\beta = 1.0$, $\gamma = 0.5$, $\tau = 2$, $\alpha = 0.5$, 1 epoch, cosine LR.

MODEL	FINE-TUNING	METHOD	SPEARMAN	PEAK PWA
QWEN2.5-0.5B-INSTRUCT	MLP-ONLY	CPO	0.00	55.04
		IPO	27.79	66.36
		SIMPO	29.50	69.65
		ALPHAPO	23.69	66.11
		GAPPO	**45.49**	**80.93**
	FULL	CPO	0.00	55.78
		IPO	31.44	67.97
		SIMPO	31.02	70.34
		ALPHAPO	24.43	67.48
		GAPPO	**46.46**	**81.73**
GEMMA-2-2B	MLP-ONLY	CPO	23.59	65.16
		IPO	34.80	79.45
		SIMPO	50.46	80.80
		ALPHAPO	53.55	79.72
		GAPPO	**57.10**	**85.26**
	FULL	CPO	24.59	69.71
		IPO	36.09	75.16
		SIMPO	57.37	82.17
		ALPHAPO	56.05	84.94
		GAPPO	**61.46**	**86.22**
MISTRAL-7B-INSTRUCT	MLP-ONLY	CPO	22.92	71.31
		IPO	51.61	80.77
		SIMPO	57.66	85.42
		ALPHAPO	57.10	84.62
		GAPPO	**58.11**	**86.06**
	FULL	CPO	11.08	72.70
		IPO	35.70	79.79
		SIMPO	49.64	85.52
		ALPHAPO	46.76	85.86
		GAPPO	**50.87**	**86.34**

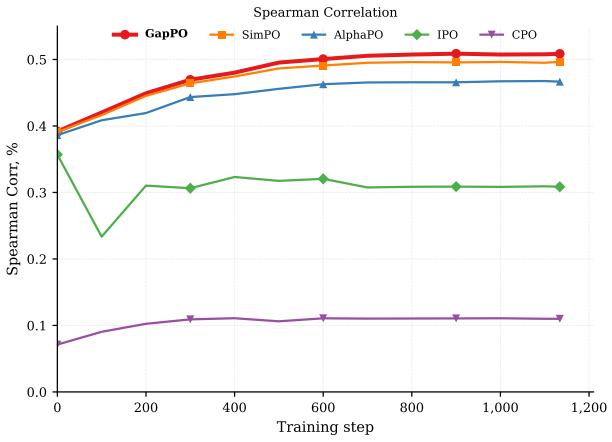

Figure 5. Spearman rank correlation between model reward and annotation scores over training steps on Mistral-7B-Instruct. GapPO's score-based weighting drives higher correlation throughout training, indicating the reward better tracks annotation quality across the full range — the prerequisite for listwise ranking.

4.5. Why Small Models Benefit Most

PWA gains diminish with model scale (Table 1), with the largest improvements at Qwen2.5-0.5B and smaller but consistent gains at Gemma-2-2B and Mistral-7B. This pattern is consistent with the GapPO mechanism. Since a small model has limited capacity, the quality of gradient signal matters more. Wasting updates on noisy low-$|\delta|$ pairs is more costly when the model has fewer parameters to absorb the noise. Larger models are more robust to training noise, so the marginal benefit of concentrating gradient on high-$|\delta|$ pairs is smaller, though still positive across all settings.

4.6. Spearman Gains Validate Reward Calibration

Spearman improvements are disproportionately large relative to PWA improvements at small scale (Table 1). This dissociation supports the core GapPO claim: the model is not merely learning to classify unambiguous pairs correctly, but developing a reward function that tracks annotation quality as a continuous signal. PWA only measures direction;

Table 2. Peak PWA for GapPO with and without weight normalization. All results use $\beta = 1.0$, $\gamma = 0.5$, $\delta = 1.0$, 1 epoch, cosine LR.

MODEL	FINE-TUNING	VARIANT	PEAK PWA (%)
GEMMA-2-2B-IT	MLP-ONLY	NO WEIGHT NORM	**81.71**
		WEIGHT NORM	80.80
	FULL	NO WEIGHT NORM	**83.26**
		WEIGHT NORM	82.35

Spearman measures calibration across the full quality spectrum. The fact that Spearman improves more than PWA suggests that $|\delta|$ weighting shapes the reward magnitude, not just its sign, which is the prerequisite for scaling to listwise NDCG evaluation where all k responses per prompt must be jointly ordered.

4.7. Consistent Method Performance Ordering

The relative ordering of most baselines is stable across scales. CPO consistently ranks last, SimPO consistently ranks second, and GapPO consistently ranks first. However, IPO and AlphaPO swap positions depending on model size. This pattern reflects a structural interaction between each method's regularization strength and model capacity. CPO systematically underperforms because it does not apply a length normalization to the log-likelihood reward, making the implicit reward strongly correlated with response length. On UltraFeedback, where chosen responses are not consistently longer than rejected ones, this conflates length with quality and produces near-chance PWA in several settings (e.g., 0.00 Spearman on Qwen2.5-0.5B). IPO avoids the CPO length bias by operating on log-probability ratios and adding a squared-deviation regularizer toward uniform preferences, which prevents reward over-optimization but also caps the reward dynamic range. AlphaPO applies a concave transformation to the reward gap, which reduces over-optimization risk but discards gradient from high-confidence pairs that are most informative when annotation quality varies widely across the dataset. At small scale (Qwen2.5-0.5B), IPO's stronger regularization is beneficial because small models are more prone to over-optimization, so IPO slightly edges out AlphaPO. At larger scales (Gemma-2-2B, Mistral-7B), the model has enough capacity to exploit the additional signal from large-gap pairs, so AlphaPO's softer regularization becomes an advantage, and it outperforms IPO. SimPO adds a length-normalization penalty and a target margin γ, which together push the reward to be more discriminative.

GapPO improves over SimPO by treating annotation scores as continuous confidence weights rather than binary labels. All four baselines assign $w = 1$ to every pair regardless of how large or small $|\delta|$ is. GapPO's $w = |\delta|$ breaks this symmetry since high-confidence pairs drive proportionally larger gradients, and the unnormalized variant additionally allows informative batches to take larger effective steps. The mechanism is additive to SimPO's length normalization and margin, so GapPO captures both the per-pair precision of SimPO and the annotation-confidence weighting that no baseline uses.

4.8. Limitations and Future Work

The current evaluation measures pairwise accuracy on held-out chosen/rejected pairs, a necessary condition for ranking quality but not a sufficient one. A model with high PWA correctly orders the specific pairs it was trained to distinguish, but this does not guarantee a globally consistent quality ordering across all responses to a prompt. Demonstrating global ranking consistency requires listwise evaluation by computing NDCG@k by running inference on all candidate responses per prompt and comparing the induced ranking to ground-truth annotation scores. We are actively developing this evaluation using the full UltraFeedback dataset (4 scored responses per prompt, yielding up to 6 pairs each), which will allow us to measure whether GapPO's pairwise gains and improved Spearman calibration translate to improved listwise NDCG.

5. Conclusion

We introduced GapPO (Gradient-Adaptive Pairwise Preference Optimization), a preference optimization method that improves pairwise ranking accuracy in LLMs by weighting each training pair by its annotation score gap. Pairwise accuracy is a direct measure of the model's usefulness as a ranker since a model that correctly orders held-out response pairs is a more reliable backbone for best-of-n sampling, reranking, and downstream RLHF. Large gaps in annotation scores indicate unambiguous comparisons that should drive larger gradients, while near-ties carry noisy labels that should be down-weighted but not discarded. GapPO implements this through linear per-sample weights proportional to $|\delta|$ and unnormalized weighting that allows batch informativeness to modulate the effective learning rate. Across four model families and two fine-tuning regimes, GapPO consistently improves peak PWA and Spearman Correlation over SimPO, CPO, IPO, and AlphaPO baselines.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, specifically improving the alignment of large language models with human preferences. GapPO is a training-time method that reweights preference pairs by annotation confidence. It does not introduce new data collection procedures, deployment pipelines, or interaction modalities. The primary societal benefit is more accurately aligned LLMs. Models whose implicit reward functions better reflect genuine human quality judgments rather than superficial features such as response length or annotation noise. More reliable alignment methods reduce the risk that deployed models reward-hack their way to high scores on proxy metrics while failing on the underlying objective.

The method could in principle be applied to align models toward harmful objectives if the training data and annotation scores encode harmful preferences. This risk is not specific to GapPO and applies equally to all preference optimization methods. GapPO does not amplify it beyond existing baselines. We evaluate exclusively on the publicly available UltraFeedback dataset, which targets helpfulness and instruction-following. We do not foresee additional ethical concerns specific to this work beyond those that are well established for the broader field of LLM alignment.

References

Azar, M. G., Rowland, M., Piot, B., Guo, D., Calandriello, D., Valko, M., and Munos, R. A general theoretical paradigm to understand learning from human feedback. *arXiv preprint arXiv:2310.12036*, 2024.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z., and Sun, M. UltraFeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36, 2024.

Geva, M., Schuster, R., Berant, J., and Levy, O. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2021.

Gupta, A., Tang, S., Mehta, S., Aggarwal, V., Garg, S., Shim, K., Lee, H., Kim, T., Kim, J., and Cho, S. AlphaPO: Reward shape matters for LLM alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Hou, Y., Zhang, J., Lin, Z., Lu, H., Xie, R., McAuley, J., and Zhao, W. X. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pp. 364–381. Springer, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2022.

Liu, T., Zhao, Y., Joshi, R., Khalman, M., Saleh, M., Liu, P. J., and Liu, J. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*, 2024.

Liu, T.-Y. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

Ma, X., Zhang, X., Pradeep, R., and Lin, J. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*, 2023.

Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Meng, Y., Xia, M., and Chen, D. SimPO: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Qin, Z., Jagerman, R., Hui, K., Zhuang, H., Wu, J., Shen, L., Liu, X., Liu, J., Metzler, D., Wang, X., et al. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1504–1518, 2024.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Sun, W., Yan, L., Ma, X., Wang, S., Ren, P., Chen, Z., Yin, D., and Ren, Z. Is ChatGPT good at search? investigating large language models as re-ranking agents. *arXiv preprint arXiv:2304.09542*, 2023.

Xu, H., Sharaf, A., Chen, Y., Tan, W., Prate, L., Murray, K., Duh, K., and Post, M. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. *arXiv preprint arXiv:2401.08417*, 2024.

Zhou, Z., Liu, J., Yang, C., Shao, J., Liu, Y., Yue, X., Ouyang, W., and Qiao, Y. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. *arXiv preprint arXiv:2310.03708*, 2023.