Low Chosen-Gradient Data Selection for Direct Preference Optimization

Anonymous ACL submission

Abstract

Direct Preference Optimization (DPO) aligns language models with human preferences using pairwise data, yet not all preference pairs contribute equally—some reduce the training loss without improving what the model actually generates at inference time. This gap stems from an asymmetry: the chosen response directly reshapes the model's high-probability decoding region, whereas the rejected response affects generation only when it overlaps with the model's current likely outputs. We propose Low-g^+, a chosen-response-centered data selection method that scores each sample by the gradient norm of the chosen response's log-likelihood, measuring how much parameter perturbation is needed to increase its likelihood. Unlike difficulty-based selection, a low g^+ does not indicate a trivial sample; it indicates a chosen response that is already geometrically compatible with the model and can be absorbed with minimal disruption to existing generation patterns. We show theoretically that smaller g^+ tightens a lower bound on the preserved probability mass in the high-probability generation region after each update. Across four models and three preference datasets, selecting only the bottom-30% g^+ samples matches or outperforms full-data DPO and existing selection baselines in most settings on AlpacaEval, MT-Bench, and Arena-Hard.

1 Introduction

Direct Preference Optimization (DPO) (Rafailov et al., 2023) directly optimizes a model toward human preferences by comparing a chosen response with a rejected response, and has been widely used due to its simple formulation and stable training (Tunstall et al., 2023; Touvron et al., 2023). However, the effectiveness of DPO heavily depends on the quality of preference data (Cui et al., 2023; Liu et al., 2023b), and different data subsets can lead to substantially different alignment outcomes (Zhou et al., 2023). Therefore, selecting more effective training samples from existing preference data is a key problem for improving both the training efficiency and alignment performance of DPO.

Existing preference data selection methods usually estimate sample value from the optimization objective of DPO, using signals such as DPO loss, reward margin, or pairwise difficulty (Liu et al., 2023a; Xia et al., 2024). These methods typically treat each preference pair as a single selection unit and focus on its pair-level optimization strength, while paying less attention to the distinct roles of the chosen and rejected responses in data selection. We argue that these two responses contribute differently to the model's actual generation distribution (Pal et al., 2024; Chen et al., 2024). As a result, selecting data solely according to the optimization strength of the whole pair may fail to accurately capture the practical value of a sample for alignment. Based on this observation, we propose a chosen-response-centered view of DPO data selection, which focuses on the positive update direction provided by the chosen response during preference alignment. We develop this view from two aspects: why the chosen response should be central to data selection, and how to select chosen responses that are better suited for stable alignment.

Why should the chosen response serve as the main basis for data selection? From the gradient form of DPO, a preference sample simultaneously increases the probability of the chosen response and decreases the probability of the rejected response; together, they define a pairwise preference update. However, during inference, the model does not uniformly explore the entire response space. Instead, decoding mainly takes place in the high-probability region of the current model distribution. In this region, the chosen and rejected responses affect generation behavior asymmetrically: the chosen response provides a constructive signal that specifies the generation pattern to be reinforced, whereas the

rejected response mainly provides a suppressive signal whose effect depends on whether it lies in the model's current high-probability error region. If the rejected response is already in a low-probability region, further suppressing it may still reduce the DPO loss but has limited impact on actual decoding behavior. Thus, in many preference samples, the chosen response more directly determines how the sample shapes the model's high-probability generation region, which motivates chosen-centered selection.

How should samples be selected based on the chosen response? DPO performs preference correction on a model that already has strong language modeling ability. Therefore, an ideal alignment update should move the model toward the preferred behavior while preserving its original capabilities and effective generation patterns as much as possible. In other words, a high-value sample is not necessarily one that induces the largest loss or the largest update, but one that can achieve preference correction with small local perturbation. To characterize this perturbation, we use the score-function gradient norm of the chosen response, denoted by g^+. This metric is not intended to measure the semantic quality of the chosen response; rather, it measures the local update cost required for the current model to absorb the chosen response. Further theoretical analysis shows that a smaller g^+ reduces the perturbation penalty in the optimization lower bound, thereby improving the effective lower bound while reducing perturbations to the original high-probability generation distribution. Based on this result, we propose Low-g^+ data selection, which prioritizes preference pairs whose chosen responses have smaller gradient norms for DPO training.

In summary, this paper proposes a chosen-response-centered framework for DPO data selection. Our main contributions are as follows:

- We reveal the mismatch between objective-level sample value and inference-time generation behavior in DPO data selection, and analyze the asymmetric roles of chosen and rejected responses in the high-probability generation region.

- We formulate preference data selection as the problem of identifying low-perturbation chosen-side updates, and use the chosen-gradient norm g^+ to characterize its local update cost.

- We explain the effectiveness of Low-g^+ selection from the perspectives of the optimization lower bound and distributional stability, showing how a smaller g^+ reduces the perturbation penalty and mitigates the impact on the original generation distribution.

- We validate the effectiveness of Low-g^+ selection across multiple experimental settings. The results show that our method identifies preference samples better suited for stable alignment and improves DPO training performance.

2 Related Work

Preference Alignment. Aligning LLMs with human preferences is central to post-training. Early RLHF methods train a reward model and optimize the policy via reinforcement learning (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022), but the multi-stage pipeline is complex and sometimes unstable (Touvron et al., 2023). DPO simplifies this by directly optimizing on pairwise preference data (Rafailov et al., 2023) and has been widely adopted (Tunstall et al., 2023). Follow-up works improve or generalize DPO through modified objectives, margins, or preference modeling (Azar et al., 2023; Amini et al., 2024; Ethayarajh et al., 2024; Hong et al., 2024; Meng et al., 2024; Hejna et al., 2023; Song et al., 2023; Liu et al., 2023a; Rosset et al., 2024; Wu et al., 2024). These focus on better optimization objectives; our work addresses a complementary problem: selecting more useful preference pairs for DPO training.

Preference Data Selection. The effectiveness of preference optimization depends strongly on data quality (Zhou et al., 2023; Liu et al., 2023b). Existing methods select or reweight pairs by objective-level signals such as preference margin, loss, reward gap, or difficulty (Amini et al., 2024; Qi et al., 2025; Das et al., 2024; Xia et al., 2024). However, most treat chosen and rejected responses as symmetric signals (Park et al., 2024; Tajwar et al., 2024; Xu et al., 2024; Feng et al., 2024). In contrast, we consider the mismatch between objective improvement and inference-time generation: a sample is valuable not only because it reduces DPO loss, but because it effectively shapes the high-probability generation region visited during decoding.

Generation Distribution and Stability. Prior analyses show that chosen response quality dominates final alignment performance (Pal et al., 2024;

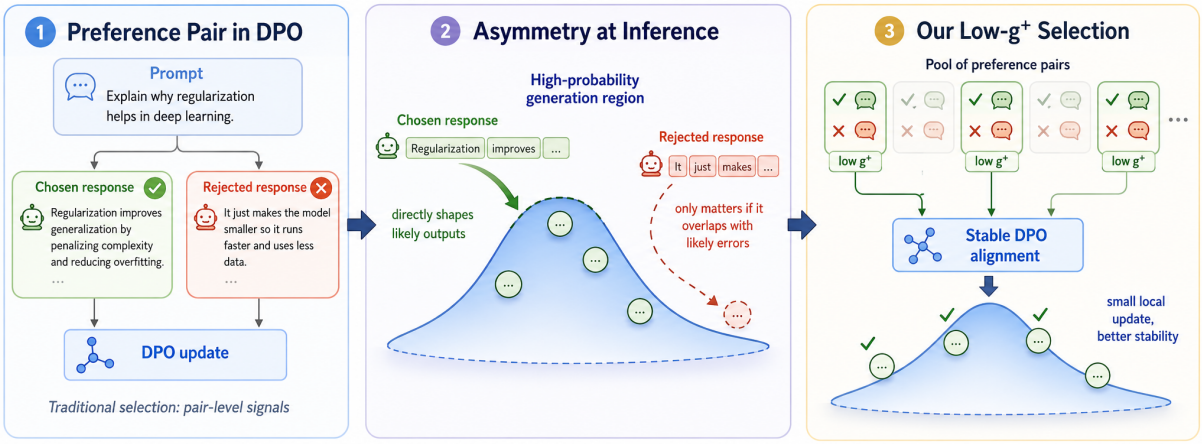

Figure 1: Overview of Low-g^+ data selection. Preference pairs are ranked by the chosen-response gradient norm $g^+(x, y^+)$, and low-g^+ pairs are selected for DPO training to reduce local update cost and preserve high-probability generation regions.

Chen et al., 2024), consistent with our chosen-response-centered view. Unlike methods using gradient or loss as direct quality indicators, we use the chosen-gradient norm g^+ to measure the local update cost for the model to absorb it. Selecting low-g^+ samples reduces distributional perturbation in the high-probability region and improves the effective lower bound of preference optimization (Yan et al., 2024; Tang et al., 2024).

3 Method

We propose Low-g^+, a chosen-response-centered data selection method for Direct Preference Optimization (DPO). Our key idea is that the value of a preference pair should not be measured only by its contribution to the DPO loss. Since inference-time decoding mainly visits a high-probability region of the current model distribution, an effective preference sample should also improve this region while introducing only a small local perturbation.

3.1 DPO as a Pairwise Gradient Update

Given a preference pair (x, y^+, y^-) and a reference model π_{ref}, DPO defines the implicit preference margin as

$$\Delta_\theta = \log \frac{\pi_\theta(y^+|x)}{\pi_{\text{ref}}(y^+|x)} - \log \frac{\pi_\theta(y^-|x)}{\pi_{\text{ref}}(y^-|x)}. \tag{1}$$

The single-sample DPO loss is

$$\ell_{\text{DPO}} = -\log \sigma(\beta \Delta_\theta), \tag{2}$$

where β is the temperature coefficient. The gradient descent direction is

$$\begin{aligned} -\nabla_\theta \ell_{\text{DPO}} &= w_\theta \big[\nabla_\theta \log \pi_\theta(y^+|x) \\ &\quad - \nabla_\theta \log \pi_\theta(y^-|x)\big], \\ w_\theta &= \beta \sigma(-\beta \Delta_\theta). \end{aligned} \tag{3}$$

This form shows that DPO treats y^+ and y^- symmetrically in the pairwise objective. However, during inference the model mainly generates from a high-probability region, where the two responses have asymmetric effects on actual generation behavior.

3.2 Asymmetry in the High-Probability Generation Region

Let $\mathcal{T}_\theta(x)$ denote the decoding-reachable high-probability generation region under the current model $\pi_\theta(\cdot|x)$:

$$\mathcal{T}_\theta(x) \subseteq \mathcal{Y}. \tag{4}$$

For example, under top-p sampling, $\mathcal{T}_\theta(x)$ contains the smallest set of responses whose cumulative probability under $\pi_\theta(\cdot|x)$ exceeds p. This region represents the responses that the model is likely to visit during inference. Therefore, a useful preference sample should not only reduce the DPO loss, but also improve or preserve the generation patterns inside $\mathcal{T}_\theta(x)$.

Let $\tilde{\pi}_\theta(\cdot|x)$ denote the model distribution restricted and renormalized to $\mathcal{T}_\theta(x)$, i.e., $\tilde{\pi}_\theta(y|x) = \pi_\theta(y|x)\,\mathbb{I}[y \in \mathcal{T}_\theta(x)]/Z_\theta(x)$ where $Z_\theta(x) = \sum_{y \in \mathcal{T}_\theta(x)} \pi_\theta(y|x)$. This distribution describes the model's effective decoding behavior restricted to the high-probability region.

From this perspective, the chosen and rejected responses play asymmetric roles. The chosen response y^+ provides a constructive signal: increasing $\pi_\theta(y^+|x)$ reinforces the token-level generation patterns it contains, directly shaping the high-probability region.

In contrast, the rejected response y^- mainly provides a suppressive signal. Suppressing y^- is most useful when it overlaps with the model's current high-probability error region. To formalize this effect, let $\mathcal{B}(y^-)$ denote the set of responses that are semantically or behaviorally equivalent to the rejected response, such as responses that express the same incorrect reasoning pattern, harmful behavior, or low-quality instruction-following mode. We define the overlap between the rejected response and the top-region behavior as

$$\begin{aligned}\rho_\theta(x, y^-) &= \tilde{\pi}_\theta(\mathcal{B}(y^-) \mid x) \\ &= \sum_{y \in \mathcal{T}_\theta(x) \cap \mathcal{B}(y^-)} \tilde{\pi}_\theta(y|x). \end{aligned} \tag{5}$$

Here, $\rho_\theta(x, y^-)$ is a conceptual quantity (not a computable selection score) that measures how much the rejected response overlaps with the model's current likely outputs. When $\rho_\theta(x, y^-)$ is small, suppressing y^- has limited effect on actual decoding behavior.

In many off-policy preference datasets, this small-overlap regime is common: the rejected response is often sampled from a weaker or earlier policy and does not lie in the current policy's high-probability region. This motivates estimating sample value primarily from the chosen side.

The effective utility of a preference pair on the high-probability generation region can be decomposed into a chosen-side shaping term $U_{\text{shape}}(x, y^+; \mathcal{T}_\theta)$ and a rejected-side suppression term modulated by $\rho_\theta(x, y^-)$. When $\rho_\theta(x, y^-)$ is small—as is common in off-policy preference datasets—the rejected-side contribution becomes negligible, and the selection value is dominated by the chosen side:

$$S(x, y^+, y^-) \approx S^+(x, y^+), \tag{6}$$

where $S^+(x, y^+)$ is the value estimated from the chosen response.

3.3 Low-g^+ as Local Update Cost

Having established the chosen response as the main basis for selection, we next define how to score it. A high-value preference sample should not necessarily be the one that induces the largest update. Since DPO is applied to a model that already has strong language modeling ability, an ideal preference update should move the model toward the chosen response while preserving its existing useful generation patterns. We therefore prefer chosen responses that can be absorbed by the current model with a small local update cost.

We measure this cost using the score-function gradient norm of the chosen response:

$$g^+(x, y^+) = \left\| \nabla_\theta \log \pi_\theta(y^+|x) \right\|_2 . \tag{7}$$

A smaller $g^+(x, y^+)$ means that increasing the likelihood of y^+ requires a smaller local parameter update. This quantity is not used to judge the semantic quality of y^+. The preference annotation already provides the quality signal by identifying y^+ as the chosen response over y^-. The role of g^+ is to quantify how costly it is for the current model to absorb the chosen response.

This score is directly related to the DPO update. Let $\delta\theta^+$ denote the chosen-side component of the DPO update induced by y^+, and let η be the learning rate. From the DPO gradient form, we have

$$\delta\theta^+ = \eta w_\theta \nabla_\theta \log \pi_\theta(y^+|x), \tag{8}$$

where $w_\theta = \beta\sigma(-\beta\Delta_\theta)$ and $0 < w_\theta \leq \beta$. Therefore,

$$\left\| \delta\theta^+ \right\|_2 \leq \eta\beta g^+(x, y^+). \tag{9}$$

Thus, under fixed training hyperparameters, g^+ directly controls an upper bound on the magnitude of the chosen-side update. Selecting samples with smaller g^+ therefore favors preference updates that can promote the chosen response with smaller local perturbation.

We further show that smaller g^+ leads to better preservation of the high-probability generation region. Under a local smoothness assumption, the perturbation to log-probabilities in $\mathcal{T}_\theta(x)$ can be bounded by a term $\epsilon(x, y^+)$ that increases monotonically with $g^+(x, y^+)$. Let $\mathcal{G}_\theta(x) \subseteq \mathcal{T}_\theta(x)$ denote the high-quality subset of the current top region. Then the normalized mass after the chosen-side update satisfies

$$\begin{aligned}&\tilde{\pi}_{\theta+\delta\theta^+}(\mathcal{G}_\theta(x) \mid x) \\ &\quad \geq \exp(-2\epsilon(x, y^+))\, \tilde{\pi}_\theta(\mathcal{G}_\theta(x) \mid x).\end{aligned} \tag{10}$$

Since $\epsilon(x, y^+)$ increases with $g^+(x, y^+)$, selecting low-g^+ samples minimizes the mass loss in the top

region. The full derivation, including the explicit form of $\epsilon(x, y^+)$ in terms of g^+, is provided in Appendix A.

3.4 Selection Rule

For each preference pair (x, y^+, y^-) in the candidate dataset $\mathcal{D}$, we compute the chosen-gradient norm $g^+(x, y^+)$. Since longer responses may naturally accumulate larger gradient norms, we use a length-normalized score for ranking:

$$\bar{g}^+(x, y^+) = \frac{1}{\sqrt{|y^+|}} \cdot \left\| \nabla_\theta \log \pi_\theta(y^+|x) \right\|_2 . \quad (11)$$

We then select the preference pairs with the smallest $\bar{g}^+$:

$$\mathcal{D}_{\text{sel}} = \text{BottomK}_{(x,y^+,y^-)\in\mathcal{D}}\, \bar{g}^+(x, y^+). \quad (12)$$

The selected subset is trained with the standard DPO objective unchanged. Low-g^+ is a data selection strategy: it does not modify the training objective, and the rejected response remains part of the DPO loss. Unless otherwise specified, Low-g^+ selection in experiments refers to ranking by the length-normalized score $\bar{g}^+$.

4 Mechanism Validation

Our method rests on the claim that y^+ and y^- play asymmetric roles in the high-probability generation region (Section 3.2). To further substantiate this claim, we design two controlled experiments that provide direct empirical evidence.

Single-Step Generation Impact. We isolate the respective effects of y^+ and y^- on model generation through a single parameter update. For each sample (x, y^+, y^-), we first greedy-decode a baseline sequence $\hat{y}$ from the SFT model, then: (1) apply one gradient-ascent step on $\log \pi_\theta(y^+|x)$ and re-decode, measuring how much the new sequence moves toward y^+; (2) restore parameters, apply one gradient-descent step on $\log \pi_\theta(y^-|x)$ and re-decode, measuring how much the new sequence moves away from y^-. To control for length effects, we only retain samples where y^+ and y^- differ in length by at most 10%. We average over 500 samples per setting.

As shown in Figure 2, y^+ ascent significantly closes the gap between model generation and y^+ across all 9 settings, while y^- descent has a negligible suppression effect. This is consistent with the probability status of y^- in the model distribution: across all settings, $\log \pi(y^-|x)$ is substantially lower than $\log \pi(y^+|x)$, indicating that y^- already lies outside the high-probability generation region. Suppressing an already low-probability sequence has little impact on the actual decoding path, whereas promoting y^+ directly reshapes the model's most likely generation trajectory.

Setting	**FDP** y^+	**FDP** y^-	Δ
Llama-3.2-3B×Sky	10.38	12.26	−1.88
Llama-3.2-3B×UF	11.89	15.46	−3.57
Llama-3.2-3B×Nec	10.66	15.24	−4.58
Llama-3.1-8B×Sky	2.85	4.79	−1.94
Llama-3.1-8B×UF	3.78	6.19	−2.41
Llama-3.1-8B×Nec	4.71	6.74	−2.03
Qwen3-4B×Sky	1.98	9.90	−7.92
Qwen3-4B×UF	3.05	12.30	−9.25
Qwen3-4B×Nec	1.79	11.63	−9.84

Table 1: Average first divergence position (FDP) after a single gradient step. Lower FDP means the generation changes earlier. y^+ ascent consistently diverges at earlier positions than y^- descent, indicating stronger influence on the autoregressive generation chain.

We further examine the *first divergence position* (FDP), defined as the earliest token position where the updated greedy sequence differs from the baseline. Table 1 reports the average FDP for each setting.

Across all settings, y^+ ascent causes the generation to diverge at earlier positions (lower FDP) than y^- descent. On Qwen3-4B, the gap is striking: y^+ diverges within the first 2–3 tokens while y^- only diverges around token 10–12. This confirms that promoting the chosen response directly perturbs the high-probability autoregressive chain from its earliest steps, whereas suppressing the rejected response has a more localized and delayed effect.

Robustness to Rejected Response Quality. To verify from an end-to-end training perspective that pair utility is dominated by y^+, we design a controlled experiment using UltraFeedback prompts with 4 ranked responses per prompt (rank 1 = best). We construct two groups of DPO datasets sharing the same prompt set: Group A fixes y^+ = rank1 and Group B fixes y^+ = rank2, with y^- drawn from rank3 or rank4 within each group. We train full DPO (no selection) on each dataset and evaluate AlpacaEval 2 pairwise win rate against SFT.

As shown in Table 2, within each group, changing y^- quality from rank3 to rank4 produces only

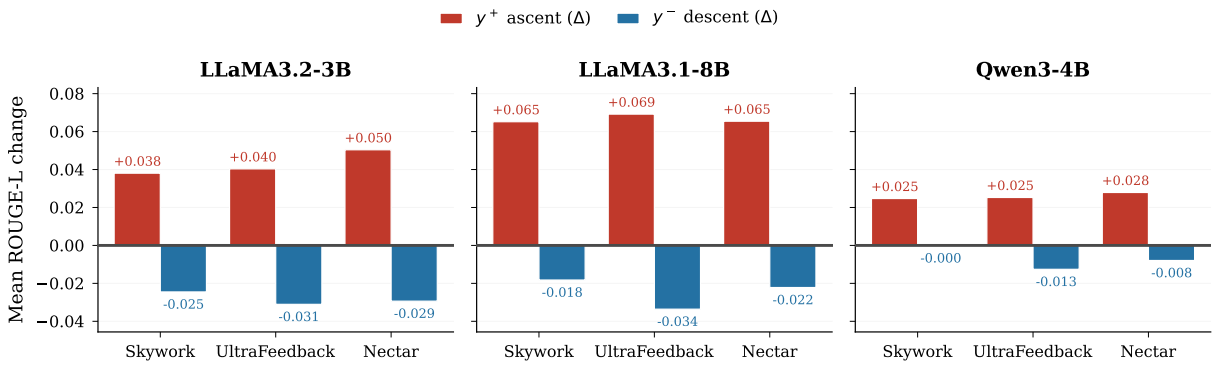

Figure 2: Mean ROUGE-L change after a single gradient step. y^+ ascent (red) substantially moves generation toward y^+, while y^- descent (blue) produces negligible suppression effect across all models and datasets.

Model	rank1 vs 3	rank1 vs 4
Llama-3.2-3B	79.1	77.3
Qwen3-4B	79.4	81.7
Model	**rank2 vs 3**	**rank2 vs 4**
Llama-3.2-3B	74.3	72.9
Qwen3-4B	70.9	74.7

Table 2: AlpacaEval 2 win rate (%) against SFT. Within each group (fixed y^+), varying y^- quality produces marginal differences; across groups, changing y^+ quality (rank1→rank2) causes a consistent 5–8 point drop.

marginal differences (Llama-3B: 79.1 vs 77.3; Qwen3-4B: 79.4 vs 81.7), while across groups, lowering y^+ quality from rank1 to rank2 leads to a consistent 5–8 point drop. This directly demonstrates that the training utility of a preference pair is primarily determined by y^+ quality rather than y^-, supporting the approximation $S(x, y^+, y^-) \approx S^+(x, y^+)$.

5 Experiments

5.1 Experimental Setup

Models. We evaluate on four base models spanning two architectures and multiple scales: Llama-3.2-3B-Base, Llama-3.1-8B-Base (Grattafiori et al., 2024), Qwen3-4B-Base, and Qwen3-8B-Base (Yang et al., 2025). All models are first supervised fine-tuned on UltraChat-200k (Ding et al., 2023), and DPO training starts from these SFT checkpoints.

Preference Datasets. We use three publicly available preference datasets: Skywork (Liu et al., 2024) (70k pairs) with high-quality human annotations, UltraFeedback (Cui et al., 2023) (57k pairs) covering diverse instruction types, and Nectar (Zhu et al., 2023) (123k pairs) with broad topic coverage.

Evaluation. We evaluate all models on AlpacaEval 2 (Dubois et al., 2024) vs SFT (pairwise win rate over 805 instructions, judged by GPT-4), MT-Bench (Zheng et al., 2023) (multi-turn quality scored 1–10 across 8 categories, judged by GPT-4), and Arena-Hard vs SFT (win rate on 500 challenging prompts, judged by GPT-4). We report the best epoch result for each experiment. Detailed training configuration and gradient scoring procedure are provided in Appendix B.

Baselines. We compare Low-g^+ against five baselines at 30% selection ratio: **Random** (uniform sampling), **Full** (100% data, no selection), **BeeS** (Deng et al., 2025) (Bayesian aggregation of external and implicit reward margins), **SelDPO** (Gao et al., 2025) (retaining easiest samples by cross-validated DPO loss), and **Difficulty** (Qi et al., 2025) (selecting the hardest samples with the smallest implicit reward margin between chosen and rejected responses).

5.2 Main Results

Table 3 presents the main results across four models and three datasets at 30% selection ratio. Low-g^+ achieves the best or second-best score in 55 out of 60 metric entries (4 models × 3 datasets × 5 metrics), ranking first 36 times—far exceeding the next-best methods SelDPO (11 firsts, 10 seconds) and BeeS (7 firsts, 12 seconds).

What stands out most is the cross-setting stability of Low-g^+. On AlpacaEval vs SFT (AE_S), the most direct measure of alignment quality, Low-g^+ ranks first or second in 11 out of 12 settings, outperforming Full by +6.3 points on average, BeeS by +4.0, and Random by +7.1. On MT-Bench, Low-g^+ achieves the highest score in 8 out of 12 settings, with an average of 6.32 compared to 6.07 for Full

Method	Skywork					UltraFeedback					Nectar				
	AE_S	AE_R	MT	AH_S	AH_R	AE_S	AE_R	MT	AH_S	AH_R	AE_S	AE_R	MT	AH_S	AH_R
Llama-3.2-3B															
Low-g^+ (Ours)	**78.0**	**7.70**	**6.21**	**55.8**	**13.8**	**85.7**	11.74	**6.31**	56.9	15.4	**83.5**	**10.31**	6.31	57.8	**15.2**
Full	69.5	6.31	6.00	55.3	13.7	81.4	10.19	6.00	54.2	15.2	78.2	9.55	6.31	53.2	14.0
SelDPO	57.0	7.08	6.07	51.9	13.2	82.7	**12.92**	6.24	**58.2**	16.2	79.6	9.40	**6.34**	54.6	13.1
BeeS	65.1	5.84	5.92	52.6	12.0	83.9	10.47	6.29	56.8	**16.7**	81.9	10.25	6.25	**58.0**	15.2
Difficulty	68.7	7.45	6.03	55.5	13.2	71.8	7.83	6.01	51.0	14.1	81.0	9.69	6.15	53.0	13.9
Random	66.4	6.20	5.88	52.3	12.9	79.7	10.86	6.12	53.2	14.4	80.4	9.80	6.12	51.2	13.0
Llama-3.1-8B															
Low-g^+ (Ours)	**81.0**	**8.61**	**6.38**	**58.5**	**16.9**	**87.8**	17.67	**6.53**	65.8	**25.2**	**84.7**	**13.79**	**6.46**	**62.0**	**22.8**
Full	67.8	8.20	6.26	58.3	16.3	83.1	13.93	6.30	63.2	23.4	84.1	12.41	6.14	61.8	22.4
SelDPO	74.3	7.70	6.04	58.2	15.8	83.6	**18.51**	6.39	**66.0**	23.2	79.2	12.17	5.98	60.9	21.4
BeeS	71.7	7.70	6.02	56.5	15.6	84.0	11.80	6.30	63.6	22.6	83.0	12.67	6.33	60.9	22.1
Difficulty	66.6	8.07	5.88	55.3	15.9	71.5	9.44	5.79	56.0	16.3	79.8	12.30	6.25	59.9	20.3
Random	65.8	7.73	6.12	55.3	15.3	78.8	8.46	6.28	59.3	21.3	81.0	11.68	6.33	59.3	21.3
Qwen3-4B															
Low-g^+ (Ours)	**68.9**	6.21	5.79	**60.7**	12.6	**88.2**	9.07	6.30	**75.5**	25.7	85.4	13.18	**6.44**	74.2	**26.2**
Full	41.5	6.30	5.72	59.2	12.4	79.3	9.21	5.92	75.4	25.2	**88.1**	**14.95**	6.22	69.0	24.2
SelDPO	64.5	6.09	5.64	56.8	10.4	85.2	**11.38**	**6.37**	72.3	**26.1**	86.1	13.67	6.28	74.1	25.4
BeeS	66.6	**6.71**	**5.89**	60.0	**13.0**	82.2	9.18	6.26	74.3	24.6	82.5	13.29	6.13	**75.9**	25.9
Difficulty	62.5	6.34	5.36	57.0	11.1	69.7	5.84	5.51	60.2	16.8	71.4	9.11	5.94	64.1	17.1
Random	67.0	5.96	4.80	58.3	11.4	77.3	7.70	6.26	72.1	23.3	81.1	10.03	6.13	68.3	22.2
Qwen3-8B															
Low-g^+ (Ours)	**60.3**	**6.58**	**5.52**	**52.8**	15.2	77.9	**9.88**	**7.42**	61.2	14.1	80.0	8.99	6.21	**66.5**	17.2
Full	59.1	5.94	5.42	51.3	**15.8**	72.4	7.38	6.33	60.2	13.9	**81.2**	**9.01**	**6.22**	65.3	17.2
SelDPO	52.3	5.73	5.34	47.2	9.5	**78.6**	9.29	7.20	**62.0**	**14.2**	75.3	7.42	6.17	63.4	15.3
BeeS	58.6	5.56	5.49	49.1	9.2	74.6	9.01	6.19	58.3	13.9	79.2	7.91	6.21	66.2	**17.4**
Difficulty	57.8	5.59	5.30	49.0	12.3	64.2	8.92	6.01	57.2	13.2	78.3	8.23	6.20	64.3	15.2
Random	56.8	5.71	5.12	48.0	10.2	68.9	5.94	5.75	59.2	12.3	73.0	7.19	6.16	62.1	16.3

Table 3: Main results across four models and three preference datasets. AE_S/AE_R: AlpacaEval 2 win rate (%) vs SFT / raw; MT: MT-Bench (1–10); AH_S/AH_R: Arena-Hard win rate (%) vs SFT / raw. All selection methods use 30% of the data. **Bold**: best; underline: second best.

and 6.11 for BeeS. On Arena-Hard vs SFT (AH_S), Low-g^+ ranks first in 7 out of 12 settings, averaging +1.8 over Full. This consistency indicates that the samples retained by Low-g^+ selection are not specific to a particular task type or model scale, but constitute a universally effective high-quality subset.

In contrast, other selection methods exhibit notable instability across settings. SelDPO performs strongly on UltraFeedback (particularly AH_S and AE_R), but fluctuates dramatically on Skywork—on Llama-3.2-3B its AE_S is only 57.0, well below Random's 66.4, and on Qwen3-8B it falls nearly 5 points below Random. BeeS is the most stable baseline overall, yet it never surpasses Low-g^+ on AE_S and shows clear degradation on Qwen3-8B $\times$ Skywork (58.6 vs Full 59.1). Difficulty ranks last in the vast majority of settings and even collapses severely on UltraFeedback (Qwen3-4B: 69.7, 7.6 points below Random), demonstrating that selecting "ambiguous" pairs by small reference log-probability gaps is not merely ineffective but actively harmful. Random selection, despite its simplicity, outperforms Difficulty in most settings, further confirming that naive difficulty-based filtering is worse than no filtering at all.

This contrast reveals a key distinction: methods based on external signals (e.g., reward margins, loss ranking) tend to overfit their heuristic assumptions on specific dataset–model combinations, resulting in poor cross-setting transferability. Low-g^+, by directly measuring the parameter perturbation required for the model to absorb each sample in its current state, provides a signal that is agnostic to the particular data distribution or model architecture, thereby maintaining consistent positive gains

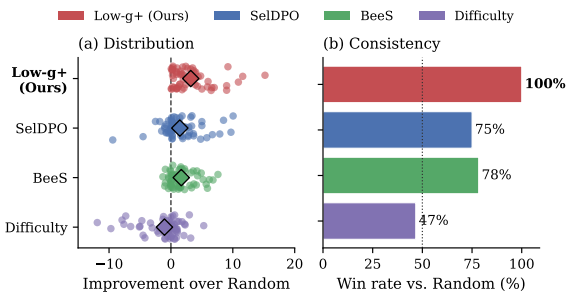

Figure 3: Stability of different selection methods across all 60 settings (4 models × 3 datasets × 5 metrics). (a) Distribution of per-setting improvement over Random; diamond markers indicate the mean. (b) Win rate: fraction of settings where each method outperforms Random.

Setting	Low-g^+	Low-γ	High-cos	Random	Low-g^-
L3B×Sky	**78.0**	70.8	66.9	66.4	59.0
L3B×UF	**85.7**	80.0	76.9	79.7	76.6
L3B×Nec	**83.5**	81.1	80.6	80.4	79.5
L8B×Sky	**81.0**	80.2	73.6	65.8	62.7
L8B×UF	**87.8**	86.6	81.1	78.8	83.3
L8B×Nec	**84.7**	82.8	81.0	81.0	82.1
Q4B×Sky	**68.9**	62.8	66.5	67.0	56.0
Q4B×UF	**88.2**	73.4	78.1	77.3	63.1

Table 4: Gradient component ablation. Let $v^+ = \nabla_\theta \log \pi_\theta(y^+|x)$ and $v^- = \nabla_\theta \log \pi_\theta(y^-|x)$. Low-$g^+$: select by $\|v^+\|^2$; Low-γ: by $\|v^+ - v^-\|^2$; High-cos: by $\cos(v^+, v^-)$; Low-g^-: by $\|v^-\|^2$. Low-g^+, Low-γ, and Low-g^- select the bottom 30%; High-cos selects the top 30%. (L3B = Llama-3.2-3B, L8B = Llama-3.1-8B, Q4B = Qwen3-4B.)

across diverse conditions. Figure 3 further visualizes this stability gap: Low-g^+ achieves the highest mean improvement over Random with the tightest variance, and wins in the largest fraction of settings among all methods.

5.3 Analysis

Validating the Chosen-Response-Centered View. To empirically validate the theoretical analysis in Section 3, we compare five gradient-based selectors that isolate different components of the DPO gradient (Table 4, 30% selection ratio, pairwise win rate against SFT judged by GPT-4).

The results reveal a striking asymmetry: Low-g^+ consistently and substantially outperforms Low-g^-, with the gap reaching 18.3 points on L8B×Skywork (81.0 vs 62.7) and 25.1 points on Q4B×UltraFeedback (88.2 vs 63.1). This directly validates our theoretical claim that the chosen response dominates the effective parameter update in the high-probability generation region. Notably, Low-g^- performs worse than random selection in most settings, confirming that the rejected response's gradient norm carries little useful signal for identifying valuable training samples. Furthermore, Low-g^+ and Low-γ (which selects by $\|g^+ - g^-\|^2$) perform comparably—Low-g^+ is slightly better on Qwen models while Low-γ edges ahead on Llama-8B. This is expected: since $\|g^+ - g^-\|^2 = \|g^+\|^2 - 2\langle g^+, g^-\rangle + \|g^-\|^2$, the expression is dominated by the $\|g^+\|^2$ term when the two gradient norms are correlated. The high-cosine selector and random selection occupy intermediate positions, further supporting that the chosen-gradient norm g^+ is the primary informative signal for data selection.

Data Efficiency. We further examine different selection ratios (20%, 50%), computational efficiency, and general knowledge preservation (MMLU/MMLU-Pro) in Appendix B.5, B.4, and B.6. Results confirm that Low-g^+ at 30% consistently matches or exceeds full-data training while using less than half the total compute, and does not degrade general capabilities.

6 Conclusion

We present a chosen-response-centered framework for DPO data selection. By analyzing the asymmetric roles of chosen and rejected responses in the high-probability generation region, we show that the chosen response dominates the effective parameter update during preference alignment, while the rejected response has limited influence on actual decoding behavior when it already lies in a low-probability region. Based on this insight, we propose Low-g^+ data selection, which prioritizes preference pairs with smaller chosen-gradient norms g^+, corresponding to lower local update cost and reduced distributional perturbation. Experiments across four models and three preference datasets demonstrate that Low-g^+ selection with only 30% of the data consistently matches or outperforms full-data training as well as existing selection baselines. Mechanism validation and gradient component ablation further provide direct empirical support for the chosen-centered view.

Limitations

Our method relies on gradient norms computed from the current policy and therefore depends on the model checkpoint used for scoring; when the

base model changes substantially, the scores need to be recomputed. Although we adopt LoRA-based proxy scoring to reduce overhead, it still requires additional forward-backward passes compared with random selection. Finally, our experiments cover models at the 3B–8B scale; validation on larger models is left for future work.

References

Afra Amini, Tim Viber, and Ryan Cotterell. 2024. Direct preference optimization with an offset. *Preprint*, arXiv:2402.10571.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Rémi Munos, and Michal Valko. 2023. A general theoretical paradigm to understand learning from human preferences. *Advances in Neural Information Processing Systems*, 36.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Angelica Chen, Sadhika Malladi, Lily H. Zhang, Xinyi Chen, and Qiuyi Zhang. 2024. Preference learning algorithms do not learn preference rankings. *Preprint*, arXiv:2405.19534.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. UltraFeedback: Boosting language models with scaled AI feedback. *Preprint*, arXiv:2310.01377.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. 2024. Active preference optimization for sample efficient RLHF. *Preprint*, arXiv:2402.10500.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. 2025. Less is more: Improving LLM alignment via preference data selection. *Preprint*, arXiv:2502.14560.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *Preprint*, arXiv:2305.14233.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model alignment as prospect theoretic optimization. *Preprint*, arXiv:2402.01306.

Duanyu Feng, Bowen Qin, Chen Huang, Zheng Zhang, and Wenqiang Lei. 2024. Towards analyzing and understanding the limitations of DPO: A theoretical perspective. *Preprint*, arXiv:2404.04626.

Chengqian Gao, Haonan Li, Liu Liu, Zeke Xie, Peilin Zhao, and Zhiqiang Xu. 2025. Principled data selection for alignment: The hidden risks of difficult examples. *Preprint*, arXiv:2502.09650.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, and Scott Niekum. 2023. Contrastive preference learning: Learning from human feedback without RL. *Preprint*, arXiv:2310.13639.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic preference optimization without reference model. *Preprint*, arXiv:2403.07691.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, and 1 others. 2024. Skywork-reward: Bag of tricks for reward modeling in LLMs. *Preprint*, arXiv:2410.18451.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J. Liu, and Jialu Liu. 2023a. Statistical rejection sampling improves preference optimization. *Preprint*, arXiv:2309.06657.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023b. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *Preprint*, arXiv:2312.15685.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. SimPO: Simple preference optimization with a reference-free reward. *Preprint*, arXiv:2405.14734.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with DPO-positive. *Preprint*, arXiv:2402.13228.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. *Preprint*, arXiv:2403.19159.

Xuan Qi, Rongwu Xu, and Zhijing Jin. 2025. Difficulty-based preference data selection by DPO implicit reward gap. *Preprint*, arXiv:2508.04149.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. 2024. Direct Nash optimization: Teaching language models to self-improve with general preferences. *Preprint*, arXiv:2404.03715.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2023. Preference ranking optimization for human alignment. *Preprint*, arXiv:2306.17492.

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

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. 2024. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. *Preprint*, arXiv:2404.14367.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Taber, Rémi Munos, Mark Rowland, Bilal Piot, and Mohammad Gheshlaghi Azar. 2024. Understanding the performance gap between online and offline alignment algorithms. *Preprint*, arXiv:2405.08448.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, and 1 others. 2023. Zephyr: Direct distillation of LM alignment. *Preprint*, arXiv:2310.16944.

Yue Wu, Zhiqing Sun, Huizhuo Yuan, Kaixuan Ji, Yiming Yang, and Quanquan Gu. 2024. Self-play preference optimization for language model alignment. *Preprint*, arXiv:2405.00675.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting influential data for targeted instruction tuning. *Preprint*, arXiv:2402.04333.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. Is DPO superior to PPO for LLM alignment? a comprehensive study. *Preprint*, arXiv:2404.10719.

Yuzi Yan, Yibo Miao, Jialian Li, Yipin Zhang, and Jian Xie. 2024. 3D-properties: Identifying challenges in DPO and charting a path forward. *Preprint*, arXiv:2406.07327.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and 1 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, and 1 others. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. LIMA: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7b: Improving LLM helpfulness & harmlessness with RLAIF. *Preprint*, arXiv:2312.09849.

A Proof of the Top-Region Mass Preservation Bound

We provide the detailed derivation of the mass preservation bound stated in Section 3.3. The goal is to show that if the log-probability perturbation is uniformly bounded on the top region, then the normalized mass assigned to any subset cannot decrease too rapidly.

Proposition 1. *Let $\mathcal{T}_\theta(x)$ be the top region and $\mathcal{G}_\theta(x) \subseteq \mathcal{T}_\theta(x)$ be any subset. Suppose that for all $y \in \mathcal{T}_\theta(x)$,*

$$|\log \pi_{\theta'}(y|x) - \log \pi_\theta(y|x)| \leq \epsilon, \tag{13}$$

where $\theta' = \theta + \delta\theta^+$. Then

$$\tilde{\pi}_{\theta'}(\mathcal{G}_\theta(x) \mid x) \geq \exp(-2\epsilon)\, \tilde{\pi}_\theta(\mathcal{G}_\theta(x) \mid x). \tag{14}$$

Proof. The uniform bound on log-probability perturbation implies that for every $y \in \mathcal{T}_\theta(x)$:

$$e^{-\epsilon} \leq \frac{\pi_{\theta'}(y|x)}{\pi_\theta(y|x)} \leq e^{\epsilon}. \tag{15}$$

Step 1: Lower bound on the unnormalized mass of $\mathcal{G}_\theta(x)$.

Summing the left inequality over $y \in \mathcal{G}_\theta(x)$:

$$\begin{aligned} &\sum_{y \in \mathcal{G}_\theta(x)} \pi_{\theta'}(y|x) \\ &\quad \geq e^{-\epsilon} \sum_{y \in \mathcal{G}_\theta(x)} \pi_\theta(y|x). \end{aligned} \tag{16}$$

Step 2: Upper bound on the unnormalized mass of $\mathcal{T}_\theta(x)$.

Summing the right inequality over $y \in \mathcal{T}_\theta(x)$:

$$\sum_{y\in\mathcal{T}_\theta(x)} \pi_{\theta'}(y|x) \leq e^{\epsilon} \sum_{y\in\mathcal{T}_\theta(x)} \pi_\theta(y|x). \tag{17}$$

Step 3: Combining via normalization.

The normalized mass under θ' is

$$\tilde{\pi}_{\theta'}(\mathcal{G}_\theta(x) \mid x) = \frac{\sum_{y\in\mathcal{G}_\theta(x)} \pi_{\theta'}(y|x)}{\sum_{y\in\mathcal{T}_\theta(x)} \pi_{\theta'}(y|x)}. \tag{18}$$

Applying the lower bound from Step 1 to the numerator and the upper bound from Step 2 to the denominator:

$$\begin{aligned}\tilde{\pi}_{\theta'}(\mathcal{G}_\theta(x) \mid x) &\geq \frac{e^{-\epsilon}\sum_{y\in\mathcal{G}_\theta(x)} \pi_\theta(y|x)}{e^{\epsilon}\sum_{y\in\mathcal{T}_\theta(x)} \pi_\theta(y|x)} \\ &= e^{-2\epsilon}\,\tilde{\pi}_\theta(\mathcal{G}_\theta(x) \mid x).\end{aligned} \tag{19}$$

This completes the proof. □

Perturbation bound derivation. We derive the uniform perturbation bound $\epsilon(x, y^+)$ used in Proposition 1. For any response $y \in \mathcal{T}_\theta(x)$, a first-order Taylor expansion gives

$$\begin{aligned}&\log \pi_{\theta+\delta\theta^+}(y|x) \\ &\quad\approx \log \pi_\theta(y|x) + \left\langle \nabla_\theta \log \pi_\theta(y|x),\, \delta\theta^+ \right\rangle.\end{aligned} \tag{20}$$

Under the local smoothness assumption, the approximation error is controlled by the second-order term. Defining $G_{\mathcal{T}} = \sup_{y\in\mathcal{T}_\theta(x)} \|\nabla_\theta \log \pi_\theta(y|x)\|_2$ and assuming $L_{\mathcal{T}}$-smoothness of the log-probability function on the top region, the worst-case log-probability change over the top region satisfies

$$\begin{aligned}&\sup_{y\in\mathcal{T}_\theta(x)} |\log \pi_{\theta+\delta\theta^+}(y|x) - \log \pi_\theta(y|x)| \\ &\quad\leq G_{\mathcal{T}}\|\delta\theta^+\|_2 + \frac{L_{\mathcal{T}}}{2}\|\delta\theta^+\|_2^2 = \epsilon(x, y^+).\end{aligned} \tag{21}$$

Combining with $\|\delta\theta^+\|_2 \leq \eta\beta\, g^+(x, y^+)$ from Section 3.3, we obtain

$$\begin{aligned}\epsilon(x, y^+) &\leq G_{\mathcal{T}}\, \eta\beta\, g^+(x, y^+) \\ &\quad + \frac{L_{\mathcal{T}}}{2}\, \eta^2\beta^2 \left(g^+(x, y^+)\right)^2,\end{aligned} \tag{22}$$

which shows that $\epsilon(x, y^+)$ increases monotonically with $g^+(x, y^+)$, yielding the bound stated in the main text.

B Detailed Experimental Setup

B.1 Training Configuration

All DPO experiments use full-parameter training (no LoRA) with $\beta = 0.1$. We train for 3 epochs using DeepSpeed ZeRO-3 on 8×A100-80GB GPUs. The effective batch size is 128 (2 per device × 8 GPUs × 8 gradient accumulation steps). We use a cosine learning rate schedule with warmup ratio 0.1. The learning rate is 2×10^{-6} for 3B and 4B models, and 5×10^{-7} for 8B models. Maximum sequence length is 2048 tokens with bf16 mixed precision. All baseline methods share the same training configuration to ensure fair comparison; only the data subset differs.

B.2 Gradient Scoring and Selection

Computing per-sample chosen-gradient norms over the full model is expensive. We adopt a lightweight proxy: we first train a LoRA adapter (rank 64, $\alpha = 128$) on the full preference dataset for 100 steps with learning rate 5×10^{-5}, then compute the length-normalized chosen-gradient norm $\bar{g}^+$ through this adapter only. This reduces the scoring cost from full-model backpropagation to adapter-only backpropagation, making it feasible to score datasets with over 100k samples on a single GPU in under 2 hours.

Specifically, for each sample (x, y^+), we compute:

$$\bar{g}^+(x, y^+) = \frac{1}{\sqrt{|y^+|}} \left\|\nabla_{\theta_{\text{LoRA}}} \log \pi_\theta(y^+|x)\right\|, \tag{23}$$

where $|y^+|$ is the number of tokens in the chosen response and θ_{LoRA} denotes the adapter parameters. The $\frac{1}{\sqrt{|y^+|}}$ normalization prevents longer responses from having systematically larger gradient norms due to summing over more token positions.

We rank all samples by $\bar{g}^+(x, y^+)$ in ascending order and select the bottom 30% for DPO training. Samples with smaller $\bar{g}^+$ can be absorbed by the model with less parameter perturbation, leading to more stable updates in the high-probability generation region.

Note that the LoRA adapter is discarded after scoring. Actual DPO training is full-parameter from the original SFT checkpoint, ensuring that the scoring procedure does not bias the final model.

Method	Selection†	Training†	Relative
Full	—	4.13×10^{18}	100%
Random	0	1.24×10^{18}	30%
Difficulty (Qi et al., 2025)	4.6×10^{17}	1.24×10^{18}	41%
BeeS (Deng et al., 2025)	4.6×10^{17}	1.24×10^{18}	41%
Low-g^+ (Ours)	6.9×10^{17}	1.24×10^{18}	47%
SelDPO (Gao et al., 2025)	1.65×10^{18}	1.24×10^{18}	70%

Table 5: Estimated total FLOPs for each method (selection + training) on Llama-3.2-3B with Skywork (70k pairs), 30% selection ratio, 3 epochs. All values are approximate, computed using the $6P$ FLOPs-per-token formula. †Estimated. "Relative" is the percentage of Full training FLOPs.

B.3 Evaluation Details

AlpacaEval 2 vs SFT. We generate responses from both the DPO model and the SFT baseline on 805 AlpacaEval 2 instructions using greedy decoding (temperature 0, max 2048 tokens). GPT-4 judges which response is better in a pairwise comparison. We report the win rate of the DPO model.

MT-Bench. We use the standard 80-question MT-Bench protocol with two-turn conversations across 8 categories (writing, roleplay, reasoning, math, coding, extraction, STEM, humanities). Responses are scored on a 1–10 scale by GPT-4.

Arena-Hard vs SFT. We generate responses on 500 challenging Arena-Hard prompts and compute pairwise win rate of the DPO model against the SFT baseline, judged by GPT-4.

For all evaluations, we report the best result across 3 training epochs. MMLU and MMLU-Pro scores are reported below to verify that data selection does not degrade general knowledge capabilities.

B.4 Computational Efficiency

We estimate the total FLOPs (selection + training) for each method using Llama-3.2-3B ($P = 3.2 \times 10^9$ parameters) on Skywork (70k pairs) as a representative example. We assume an average sequence length of $T = 512$ tokens and 3 training epochs, with the standard approximation of $6P$ FLOPs per token for a forward-backward pass.

As shown in Table 5, the estimated selection overhead of Low-g^+ (one forward-backward pass on the chosen response for all 70k samples) is approximately 6.9×10^{17} FLOPs, bringing the total to roughly 47% of full training. This is slightly higher than Random (30%, no selection cost), Difficulty (Qi et al., 2025) and BeeS (Deng et al., 2025) (41%, requiring only reference-model forward passes), but substantially lower than SelDPO (Gao et al., 2025) (70%, which requires K-fold cross-validated DPO training on the full dataset to compute loss rankings). Given that Low-g^+ consistently outperforms all baselines in alignment quality (Table 3), it achieves the best performance-efficiency trade-off among all compared methods.

B.5 Results at Other Selection Ratios

Tables 6 and 7 report results at 20% and 50% selection ratios on three representative models, complementing the 30% results in the main text.

B.6 General Knowledge Preservation

To verify that data selection does not degrade general knowledge capabilities, we report MMLU and MMLU-Pro scores for all methods across the three selection ratios. As shown in Tables 8–10, Low-g^+ maintains general knowledge scores comparable to the Full baseline in most settings, suggesting that the proposed selection strategy improves alignment quality without substantially degrading broad knowledge capabilities.

Method	Skywork					UltraFeedback					Nectar				
	AE_S	AE_R	MT	AH_S	AH_R	AE_S	AE_R	MT	AH_S	AH_R	AE_S	AE_R	MT	AH_S	AH_R
Llama-3.2-3B															
Low-g^+	66.6	6.34	5.95	52.5	11.5	79.0	9.58	5.74	**54.9**	13.8	**83.5**	**11.80**	**6.31**	56.3	**16.2**
Full	**69.5**	6.31	**6.00**	**55.3**	**13.7**	**81.4**	**10.19**	6.00	54.2	15.2	78.2	9.55	6.31	53.2	14.0
SelDPO	52.1	5.59	5.97	50.2	10.2	52.1	5.59	5.88	50.2	10.2	75.4	6.89	6.13	**56.5**	12.9
BeeS	61.1	5.59	5.91	49.8	11.2	78.4	9.57	**6.31**	54.1	**15.8**	80.8	9.81	6.18	55.5	13.9
Difficulty	64.5	**6.46**	5.61	53.8	11.7	57.6	5.71	5.55	48.5	13.3	77.6	8.32	5.91	52.6	12.3
Random	65.3	6.21	5.96	53.5	12.2	76.0	7.97	5.97	49.1	12.3	80.9	8.94	6.28	51.4	12.3
Llama-3.1-8B															
Low-g^+	**71.8**	**8.57**	5.94	**58.6**	**17.8**	82.4	11.07	6.27	62.6	21.9	82.7	**13.31**	**6.39**	**63.5**	22.8
Full	67.8	8.20	**6.26**	58.3	16.3	**83.1**	13.93	**6.30**	63.2	**23.4**	**84.1**	12.41	6.14	61.8	22.4
SelDPO	57.4	5.84	5.69	50.2	16.1	82.0	**17.14**	6.25	**64.5**	22.4	80.0	12.17	6.16	60.4	19.7
BeeS	57.8	6.06	5.84	50.8	14.0	81.7	11.18	6.15	62.2	20.5	82.5	10.56	6.36	63.1	**23.2**
Difficulty	66.0	7.33	5.95	55.1	15.3	62.4	7.08	5.66	58.0	15.7	72.3	8.82	6.06	56.9	16.7
Random	64.8	6.84	5.76	55.2	14.2	73.4	9.31	5.80	57.3	17.4	76.0	7.97	5.97	59.2	12.3
Qwen3-4B															
Low-g^+	**65.6**	5.71	**5.74**	59.1	**12.4**	73.7	7.96	5.74	64.5	23.5	78.9	11.55	6.28	68.7	23.2
Full	41.5	**6.30**	5.72	**59.2**	12.4	**79.3**	**9.21**	5.92	**75.4**	**25.2**	**88.1**	**14.95**	6.22	**69.0**	**24.2**
SelDPO	51.8	4.22	5.15	48.5	9.1	72.4	6.83	**5.96**	65.9	18.1	77.6	9.51	**6.29**	62.3	21.3
BeeS	63.0	5.34	4.67	54.7	11.2	74.0	7.94	5.91	62.1	22.0	75.3	12.30	6.09	64.2	22.4
Difficulty	60.5	5.34	5.38	56.7	12.4	60.4	4.60	5.29	53.8	11.1	72.0	7.86	5.95	62.6	17.8
Random	63.6	5.87	5.49	51.4	11.3	70.8	6.71	5.78	60.2	18.9	77.5	9.28	6.11	61.2	21.0

Table 6: Results at 20% selection ratio. Notation follows Table 3. **Bold**: best; underline: second best.

Method	Skywork					UltraFeedback					Nectar				
	AE_S	AE_R	MT	AH_S	AH_R	AE_S	AE_R	MT	AH_S	AH_R	AE_S	AE_R	MT	AH_S	AH_R
Llama-3.2-3B															
Low-g^+	**78.0**	**7.20**	**6.39**	55.7	13.8	84.3	11.68	6.29	59.0	**17.2**	**82.3**	**10.24**	**6.42**	54.8	14.2
Full	69.5	6.31	6.00	55.3	13.7	81.4	10.19	6.00	54.2	15.2	78.2	9.55	6.31	53.2	14.0
SelDPO	65.1	7.20	5.93	54.9	12.7	**86.1**	**12.41**	**6.36**	**59.8**	17.2	78.3	8.94	6.09	54.2	14.4
BeeS	64.2	6.83	6.06	53.7	13.2	85.3	10.93	6.34	59.5	16.6	81.0	9.75	6.22	**55.5**	**14.9**
Difficulty	73.7	6.83	5.91	**57.7**	**14.1**	78.6	9.81	6.21	55.2	14.1	80.9	10.19	6.31	53.5	13.2
Random	69.9	6.58	6.15	52.3	12.3	81.5	6.83	6.21	57.3	14.2	81.7	9.37	6.27	50.3	13.2
Llama-3.1-8B															
Low-g^+	**78.8**	**9.57**	6.24	61.5	18.7	**87.8**	13.87	6.14	64.5	22.6	81.7	**12.92**	6.14	**64.5**	21.2
Full	67.8	8.20	**6.26**	58.3	16.3	83.1	13.93	**6.30**	63.2	23.4	84.1	12.41	6.14	61.8	22.4
SelDPO	75.3	9.07	6.03	**61.7**	**20.1**	86.2	**20.00**	6.12	**64.6**	**24.0**	76.6	10.81	5.74	61.3	19.4
BeeS	64.5	7.20	6.01	57.4	16.6	86.1	12.67	6.19	64.2	23.9	**84.4**	12.17	6.29	63.4	**23.6**
Difficulty	69.7	8.00	6.26	58.6	16.3	78.6	10.56	6.06	59.6	19.1	82.2	12.30	**6.44**	61.4	21.2
Random	68.6	8.07	6.05	55.4	15.3	82.2	10.19	6.19	57.9	20.4	79.2	10.32	5.89	58.2	19.3
Qwen3-4B															
Low-g^+	**78.4**	**7.93**	**5.82**	67.1	15.4	**88.2**	10.56	6.14	75.8	22.2	85.4	13.19	5.94	67.3	19.6
Full	41.5	6.30	5.72	59.2	12.4	79.3	9.21	5.92	75.4	25.2	**88.1**	14.95	6.22	69.0	24.2
SelDPO	76.0	6.58	5.49	**68.2**	**17.7**	84.1	**23.85**	**6.67**	**76.8**	21.7	81.7	**16.52**	6.61	**76.3**	**28.1**
BeeS	60.5	4.72	5.49	57.0	13.0	86.6	13.04	6.49	76.1	**25.8**	82.9	14.77	**6.66**	62.3	18.4
Difficulty	63.3	6.34	5.50	60.7	13.4	76.5	8.97	6.19	64.5	19.8	71.5	8.45	5.86	64.5	18.9
Random	61.8	5.64	5.17	55.2	12.5	74.3	8.46	5.99	69.2	20.1	75.0	9.23	6.14	52.3	13.2

Table 7: Results at 50% selection ratio. Notation follows Table 3. **Bold**: best; underline: second best.

Method	Skywork		UltraFeedback		Nectar	
	MMLU	MMLU-Pro	MMLU	MMLU-Pro	MMLU	MMLU-Pro
Llama-3.2-3B						
Low-g^+	57.4	28.3	55.7	27.4	55.0	27.1
Full	57.0	28.4	58.1	29.1	50.8	26.4
SelDPO	55.9	25.9	55.9	25.9	42.6	19.9
BeeS	55.9	26.4	57.8	28.7	57.7	28.7
Difficulty	57.6	29.3	54.2	27.2	58.0	28.9
Random	57.7	29.2	56.9	28.5	53.9	27.7
Llama-3.1-8B						
Low-g^+	60.9	29.5	60.5	28.8	56.6	27.9
Full	62.5	29.5	61.9	30.5	55.0	23.1
SelDPO	41.1	18.6	29.1	14.2	49.6	23.4
BeeS	59.7	28.0	61.8	30.2	60.1	28.4
Difficulty	62.0	29.5	60.4	28.9	59.7	27.1
Random	61.9	29.9	61.0	29.7	56.9	25.5
Qwen3-4B						
Low-g^+	66.2	34.8	67.1	34.7	66.1	34.4
Full	67.7	35.5	66.5	35.8	66.0	36.3
SelDPO	66.0	33.0	64.2	32.1	60.1	28.3
BeeS	65.9	33.5	68.0	35.7	65.2	34.4
Difficulty	67.5	36.6	66.9	35.2	64.0	35.6
Random	67.4	36.0	66.5	35.3	62.2	33.7

Table 8: MMLU and MMLU-Pro scores (%) at 20% selection ratio.

Method	Skywork		UltraFeedback		Nectar	
	MMLU	MMLU-Pro	MMLU	MMLU-Pro	MMLU	MMLU-Pro
Llama-3.2-3B						
Low-g^+	56.8	27.5	58.2	28.5	55.7	27.1
Full	57.0	28.4	58.1	29.1	50.8	26.4
SelDPO	56.3	25.5	51.8	25.2	42.0	19.8
BeeS	57.0	27.7	57.5	28.2	56.6	27.5
Difficulty	57.6	29.3	58.1	28.9	55.4	28.1
Random	57.4	28.3	57.9	29.4	54.1	27.1
Llama-3.1-8B						
Low-g^+	61.3	29.8	61.2	30.6	56.4	25.0
Full	62.5	29.5	61.9	30.5	55.0	23.1
SelDPO	48.4	21.5	61.1	28.3	50.5	22.7
BeeS	61.0	28.9	62.7	30.0	56.9	24.5
Difficulty	62.3	29.6	61.9	29.1	55.8	24.4
Random	62.1	29.9	61.7	29.7	45.0	20.3
Qwen3-4B						
Low-g^+	66.5	34.9	67.0	34.0	65.2	34.2
Full	67.7	35.5	66.5	35.8	66.0	36.3
SelDPO	64.7	31.3	60.8	27.1	53.9	26.0
BeeS	66.4	34.0	68.0	35.7	64.9	34.5
Difficulty	67.6	37.0	67.2	36.7	64.1	35.6
Random	67.6	37.0	66.9	35.8	64.5	34.7
Qwen3-8B						
Low-g^+	70.8	38.8	71.2	39.5	70.6	37.0
Full	71.3	39.5	71.5	38.5	70.2	37.6
SelDPO	71.1	39.5	72.3	39.0	71.2	37.2
BeeS	70.8	38.0	72.0	39.4	70.2	37.1
Difficulty	71.4	39.4	71.3	39.8	72.0	34.8
Random	71.1	39.1	71.4	39.8	71.2	34.6

Table 9: MMLU and MMLU-Pro scores (%) at 30% selection ratio.

Method	Skywork		UltraFeedback		Nectar	
	MMLU	MMLU-Pro	MMLU	MMLU-Pro	MMLU	MMLU-Pro
Llama-3.2-3B						
Low-g^+	55.9	27.5	49.2	23.3	51.1	28.1
Full	57.0	28.4	58.1	29.1	50.8	26.4
SelDPO	54.3	25.8	50.4	23.1	53.4	26.1
BeeS	57.2	27.6	58.1	28.3	51.8	26.3
Difficulty	57.6	28.8	58.1	29.5	52.1	27.2
Random	57.6	28.9	57.6	29.2	53.5	28.1
Llama-3.1-8B						
Low-g^+	61.7	29.7	56.3	25.6	51.7	22.7
Full	62.5	29.5	61.9	30.5	55.0	23.1
SelDPO	59.2	26.4	57.7	26.2	45.5	19.2
BeeS	61.4	29.0	63.1	30.9	60.9	29.0
Difficulty	62.8	29.8	62.1	30.0	59.4	28.1
Random	62.7	29.6	61.8	30.5	60.1	27.7
Qwen3-4B						
Low-g^+	66.8	34.8	66.7	33.9	62.3	32.9
Full	67.7	35.5	66.5	35.8	66.0	36.3
SelDPO	67.4	35.8	58.7	29.4	51.8	27.7
BeeS	67.2	35.4	68.0	37.1	67.8	37.7
Difficulty	67.6	37.5	67.0	36.1	65.8	37.3
Random	67.5	37.4	65.0	34.4	67.3	36.6

Table 10: MMLU and MMLU-Pro scores (%) at 50% selection ratio.