LESS DRIFT, BETTER ALIGNMENT: SELECTING THE FITTEST PREFERENCE DATA FOR OFFLINE DPO

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline Direct Preference Optimization (DPO) relies on pre-collected preference data, but mismatches between such data and the current policy can cause large local perturbations, ultimately leading to excessive policy drift. To address this issue, we seek to identify preference pairs that induce smaller local perturbations to the current policy when trained on. We study computable proxies for this local update cost and identify a generation-level asymmetry between chosen and rejected responses. Controlled generation and gradient analyses show that chosen-side updates affect decoding more directly and reliably, whereas rejected-side effects depend on their overlap with the current policy's high-probability generation region. The chosen-gradient norm therefore provides a better proxy for local update cost. We further show that this cost is only weakly coupled with preference quality and semantic coverage, so low-cost subsets need not sacrifice either. Motivated by these results, we propose Low-g^+, which ranks complete preference pairs by a length-normalized chosen-gradient score and selects the lowest-scoring subset under a fixed data budget before standard DPO training. Under local smoothness, lower cost tightens the upper bound on log-probability perturbation and raises the lower bound on retained probability mass for high-quality responses in the generation region. Across four models, three datasets, and five metrics, Low-g^+ ranks in the top two in 55 of 60 settings.

1 INTRODUCTION

Offline Direct Preference Optimization (DPO) (Rafailov et al., 2023; Tunstall et al., 2023) trains the current policy on pre-collected preference data. Because the responses may have been generated by historical or unknown behavior policies, their distribution can differ from the current policy's sampling distribution, leaving some preference pairs mismatched with the current policy (Cui et al., 2024; Tang et al., 2024; Tajwar et al., 2024; Xiong et al., 2024). Training indiscriminately on such pairs can induce large local policy perturbations and ultimately lead to excessive policy drift. Iterative DPO can mitigate this mismatch through on-policy sampling (Xiong et al., 2024; Yuan et al., 2024), but requires additional generation and labeling (Ouyang et al., 2022; Bai et al., 2022). Rather than discard valuable preference supervision in existing large-scale offline corpora, we focus on selecting samples that are compatible with the current policy while retaining informative preference signals (Zhou et al., 2023; Liu et al., 2023; Xia et al., 2024).

We therefore adopt a stability-first principle: rather than use every pair, prefer samples that the current policy can absorb with smaller local updates. At the sample level, the gradient norm provides a computable first-order proxy for this update cost. DPO's pairwise structure, however, introduces an ambiguity. Each pair yields a likelihood-increasing component for the chosen response y^+, a likelihood-decreasing component for the rejected response y^-, and their full DPO gradient. Although the two terms enter the objective symmetrically with opposite signs, these three gradient measures need not be equally informative about perturbations to generation behavior. Using these gradient measures without distinguishing their roles in generation behavior can therefore lead to suboptimal data rankings.

The distinction is behavioral. The chosen response defines a pattern to reinforce, so increasing its likelihood can directly alter what the policy generates. In contrast, decreasing the likelihood of a rejected response matters only when the corresponding undesirable pattern overlaps with the policy's

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

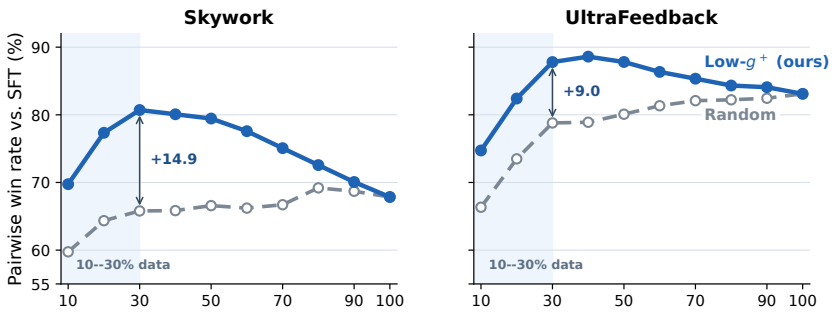

Figure 1: **Data scaling with Llama-3.1-8B.** We compare Low-g^+ with Random selection across data budgets on Skywork and UltraFeedback. Low-g^+ achieves higher AlpacaEval 2 win rates against SFT at every subsampling ratio; arrows highlight the gains at the 30% main setting.

high-probability generation region. Under offline mismatch, a rejected response may already lie in a low-probability region; further suppressing it can improve the DPO objective with little effect on decoding. Controlled one-step updates, first-divergence-position (FDP) analysis, and gradient-component ablations support this asymmetry. Among the candidate measures, the chosen-gradient norm g^+ is therefore the most direct proxy for the local update cost of perturbing current generation behavior. We further show that this cost is only weakly associated with preference quality and does not systematically reduce semantic coverage, supporting lower-cost selection without an inherent loss of supervision quality or diversity.

Building on these findings, we propose Low-g^+. Given a fixed data budget, it ranks complete preference pairs by the length-normalized chosen-gradient score $\bar{g}^+$ and selects the lowest-scoring subset. The selected pairs are then trained with the standard DPO objective, so rejected responses remain part of every update. Under local smoothness, lower chosen-side update cost tightens an upper bound on log-probability perturbation within the high-probability generation region and raises a lower bound on the normalized post-update mass assigned to its high-quality subset. Thus, Low-g^+ improves the local stability bound without relying on a systematic reduction in preference quality or diversity. Across four base models and three offline preference datasets, Low-g^+ ranks first or second in 55 of 60 model–dataset–metric combinations at a fixed 30% budget, including 36 first-place results. Figure 1 previews the same advantage across budgets: on Llama-3.1-8B, Low-g^+ exceeds Random selection from 10% through 90% of each corpus. Our contributions are threefold:

1. We identify and validate a generation-level asymmetry in DPO: chosen-side updates affect decoding more directly and reliably, whereas rejected-side effects depend on overlap with the current policy's high-probability generation region. To our knowledge, this is the first systematic comparison of isolated chosen-side and rejected-side updates on decoded outputs.
2. Based on this finding, we propose Low-g^+, which ranks complete preference pairs by a length-normalized chosen-gradient score while retaining standard DPO, and derive a local stability bound.
3. Across four models, three datasets, and five metrics, Low-g^+ ranks in the top two in 55 of 60 settings and remains effective across data budgets.

2 Related Work

Direct Preference Optimization. Reinforcement Learning from Human Feedback (RLHF) typically learns a reward model before optimizing the policy with reinforcement learning, making the pipeline complex and potentially unstable (Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022). DPO instead optimizes the policy directly from pairwise preferences (Rafailov et al., 2023; Tunstall et al., 2023). Subsequent work improves this optimization by modifying the objective, margin, reference formulation, or preference model (Azar et al., 2023; Amini et al., 2024; Ethayarajh et al., 2024; Hong et al., 2024; Meng et al., 2024; Hejna et al., 2023). Our work is complementary: we retain standard DPO and ask which pairs should be used when a fixed offline corpus is mismatched with the current policy.

Preference Data Selection. Alignment performance is highly sensitive to data quality and composition (Zhou et al., 2023; Liu et al., 2023). For a fixed preference corpus, existing methods select

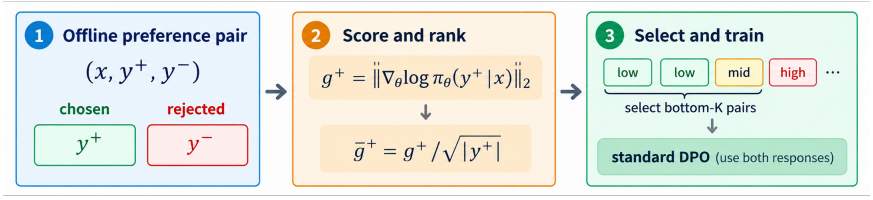

Figure 2: Low-g^+ selects complete preference pairs by the length-normalized chosen-side gradient score $\bar{g}^+$. The rejected response is excluded from ranking but retained for standard DPO training.

or reweight pairs using reward margins, implicit rewards, losses, difficulty, or alignment potential (Qi et al., 2025; Deng et al., 2025; Gao et al., 2025; Huang et al., 2025); broader data-selection work also uses gradients or influence estimates (Xia et al., 2024). These criteria primarily estimate sample quality, informativeness, difficulty, or alignment value. Recent analysis further shows that chosen-response quality can dominate rejected-response quality in DPO (Pan et al., 2025). However, prior selection methods do not explicitly treat the perturbation induced in the current policy as a sample-level cost, nor do they distinguish how chosen-side, rejected-side, and full gradients relate to generation behavior. Low-g^+ addresses this gap by separating local update cost from preference quality and deriving a side-aware selection signal from their behavioral asymmetry.

Offline Shift and Policy Stability. Offline preferences are often produced by earlier or unknown behavior policies and may therefore be mismatched with the current policy. Prior work documents the gap between online and offline alignment and the benefits of on-policy data (Tang et al., 2024; Tajwar et al., 2024). Active preference optimization adaptively acquires comparisons for uncertain contexts (Das et al., 2024); iterative preference learning and self-rewarding methods refresh data through current-policy sampling (Xiong et al., 2024; Yuan et al., 2024), while InCo-DPO mixes on- and off-policy data to balance distribution shift and quality (Wang et al., 2025). Although these approaches provide fresher supervision, they require new generations or annotations. Other work analyzes DPO failures involving length bias, preference rankings, and objective mismatch (Park et al., 2024; Chen et al., 2024a; Feng et al., 2024; Yan et al., 2024); DPO-Positive strengthens the chosen-side term by modifying the objective (Pal et al., 2024). In contrast, we reuse the fixed offline corpus, preserve the standard DPO objective, and control local policy perturbation by selecting complete pairs according to chosen-side update cost.

3 Problem Formulation

Consider a pre-collected offline preference dataset

$$\mathcal{D}_{\text{off}} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^N, \tag{1}$$

where responses were generated by a historical, external, or unknown behavior policy μ. The trainable policy π_θ is initialized from π_0, whose frozen copy serves as the reference policy. Because the data are not resampled as the policy evolves, $\mu(\cdot \mid x)$ and $\pi_0(\cdot \mid x)$ can place mass on different high-probability generation regions. Offline preference pairs therefore vary in their compatibility with the current policy. The central challenge is to identify pairs that provide useful preference correction while keeping policy perturbations controlled. Figure 2 previews our resulting selection method, Low-g^+.

For a fixed prompt x, the KL-regularized objective underlying DPO balances preference gain against policy-deviation cost:

$$\max_\pi \underbrace{\mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)]}_{\text{preference gain}} - \underbrace{\beta D_{\text{KL}}\big(\pi(\cdot \mid x) \,\|\, \pi_0(\cdot \mid x)\big)}_{\text{policy-deviation cost}}. \tag{2}$$

Its ideal non-parametric optimum satisfies $\pi_r(y \mid x) \propto \pi_0(y \mid x) e^{r(x,y)/\beta}$. Let $R_x := \sup_y r(x,y) - \inf_y r(x,y)$ denote the reward range over the support of $\pi_0(\cdot \mid x)$, and assume $R_x \le R$. Then, for any set A of high-quality responses,

$$\pi_r(A \mid x) \ge e^{-R/\beta} \pi_0(A \mid x). \tag{3}$$

This inequality lower-bounds the probability mass retained by the optimized policy on a high-quality response set. It shows that, provided preference supervision remains effective, tighter control of policy deviation can raise this retention bound. Motivated by this observation, we consider sample-level local update cost in finite-step DPO and show in our method section that reducing this cost raises the corresponding local retention bound. To determine how to measure such cost, the next section analyzes the gradient components of DPO and their effects on generation behavior. Appendix B provides the full derivation.

4 Mechanism Analysis

4.1 From Pairwise Objective to Generation Behavior

For a preference pair (x, y^+, y^-), the single-sample DPO loss relative to the reference policy π_0 is

$$\ell_{\mathrm{DPO}} = -\log \sigma\left(\beta \left[\log \frac{\pi_\theta(y^+ \mid x)}{\pi_0(y^+ \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_0(y^- \mid x)}\right]\right). \tag{4}$$

The descent direction is

$$-\nabla_\theta \ell_{\mathrm{DPO}} = w_\theta \left[\nabla_\theta \log \pi_\theta(y^+ \mid x) - \nabla_\theta \log \pi_\theta(y^- \mid x)\right]. \tag{5}$$

Here, $w_\theta > 0$ is a sample-dependent scalar determined by the DPO logit. The update contains a likelihood-increasing component for the chosen response and a likelihood-decreasing component for the rejected response, yielding three candidate signals: the chosen-side component, the rejected-side component, and the full gradient.

Although the chosen-side and rejected-side components are formally symmetric in the DPO gradient, they play different roles in generation. A chosen-side update specifies the target response y^+ toward which probability mass should move. A rejected-side update only specifies the response y^- to suppress and does not determine how the removed probability mass is redistributed among other responses.

Autoregressive decoding is governed mainly by high-probability generation trajectories under the current policy. The effect of a rejected-side update on actual generation therefore depends on whether y^- or similar undesirable patterns still have appreciable probability under the current policy. If these patterns are already rarely generated, further suppressing them can improve the DPO pairwise objective while leaving decoded outputs nearly unchanged. Offline mismatch makes this case more common. By contrast, a chosen-side update provides a concrete target trajectory and can therefore influence subsequent token generation more directly.

This distinction leads to a testable hypothesis: compared with the rejected-side gradient, the chosen-side gradient more directly and consistently characterizes local policy perturbations relevant to generation. This asymmetry concerns only the gradient signal used for data selection and does not change the DPO training objective. Each selected sample retains the complete pair (x, y^+, y^-) and is trained with standard DPO. The next subsection tests this hypothesis through controlled one-step updates.

4.2 Controlled Validation of Generation Asymmetry

We test this hypothesis with controlled one-step updates. For each preference pair, we greedily decode a baseline response from the current policy, then apply either a likelihood-increasing chosen-side update or a likelihood-decreasing rejected-side update from the identical model state, using the same learning rate, before decoding again. We use changes in standard ROUGE-L similarity (Lin, 2004) to quantify how the output shifts relative to y^+ and y^-. We also measure the *first divergence position* (FDP), the earliest token at which an updated output differs from the baseline. A smaller FDP means that the update changes the autoregressive trajectory earlier and can affect more subsequent tokens. Appendix C gives the data construction, length control, and update details.

Across model–dataset settings, chosen-side updates consistently move generated outputs toward y^+, whereas rejected-side updates produce only limited suppression of y^- under the same protocol (Figure 3). Chosen-side updates also yield smaller FDP values (Table 1), showing that they alter the autoregressive trajectory earlier. The current policy generally assigns lower likelihood to y^- than to

Table 1: Per-model FDP after controlled one-step updates (left of the rule; smaller FDP means an earlier change in the autoregressive generation path) and AlpacaEval 2 win rate (%) against SFT under varying response quality (right of the rule). rivj denotes a rank-i chosen response paired with a rank-j rejected response.

	Skywork			UltraFeedback			Nectar			AlpacaEval win %			
Model	y^+	y^-	Δ	y^+	y^-	Δ	y^+	y^-	Δ	r1v3	r1v4	r2v3	r2v4
Llama-3.2-3B	10.38	12.26	−1.88	11.89	15.46	−3.57	10.66	15.24	−4.58	79.1	77.3	74.3	72.9
Llama-3.1-8B	2.85	4.79	−1.94	3.78	6.19	−2.41	4.71	6.74	−2.03	–	–	–	–
Qwen3-4B	1.98	9.90	−7.92	3.05	12.30	−9.25	1.79	11.63	−9.84	79.4	81.7	70.9	74.7

y^+, consistent with the generation-region overlap analysis above: when a rejected response already lies in a low-probability region, further suppression has little effect on decoding.

We also test how the quality of each response affects standard DPO training. Holding the prompt and chosen response fixed while replacing the rejected response with alternatives of different quality causes only small performance changes. Lowering chosen-response quality, in contrast, produces a larger and more consistent drop (Table 1).

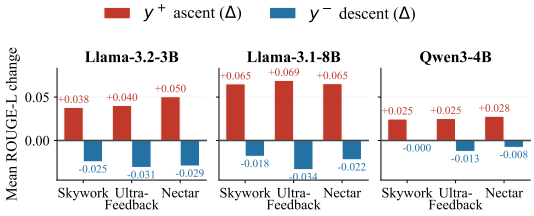

Figure 3: Generation changes after controlled one-step chosen-side and rejected-side updates, measured by mean ROUGE-L change.

This does not make rejected responses irrelevant to training. Rather, in the offline settings studied here, generation behavior and end-to-end performance are more sensitive to chosen-side information. Together, these results support our generation-level asymmetry hypothesis and motivate the chosen-side gradient as a measure of stability-related local update cost.

4.3 Update Cost and Data Attributes

The mechanism analysis and controlled validation above show that chosen-side updates are more direct and reliable indicators of local perturbations to generation. We therefore use the chosen-gradient norm g^+ as the chosen-side update-cost measure and its length-normalized form $\bar{g}^+$ as the ranking score:

$$g^+ = \left\|\nabla_\theta \log \pi_\theta(y^+ \mid x)\right\|_2, \qquad \bar{g}^+ = g^+ / \sqrt{|y^+|}. \tag{6}$$

We analyze Llama-3.1-8B (L8B) and Qwen3-8B (Q8B) on Skywork (Sky) and UltraFeedback (UF). Here, g^+ is the norm of the chosen-side likelihood gradient and serves as a proxy for local update cost; it does not measure semantic quality. Because raw gradient norm also grows with response length, normalization reduces length's influence on ranking. Figure 4(a) confirms this effect: the length correlation falls from 0.63–0.89 for g^+ to −0.12–0.06 for $\bar{g}^+$. All subsequent analyses therefore use $\bar{g}^+$.

Low update cost is useful for selection only if it does not systematically weaken preference supervision. We therefore assess chosen quality, rejected quality, and preference margin with an independent external reward model, and compute their Spearman correlations with $\bar{g}^+$. This reward model is used only for diagnosis; it never affects ranking or DPO training. Across all three measures and four settings, correlations lie between −0.16 and 0.11 (Figure 4(b)). Thus, $\bar{g}^+$ is not a simple proxy for reward-model quality, but a complementary signal of local update cost.

At the same data budget, we compare the semantic diversity of Low-g^+, Random, and High-g^+ subsets by clustering embeddings of concatenated instruction–chosen-response pairs. We report cluster coverage and cluster-distribution entropy. Low-g^+ remains comparable to Random on both measures (Figure 4(c)), rather than concentrating on a small set of homogeneous patterns. Together, these analyses show that lower chosen-side update cost is not obtained by systematically favoring shorter responses, lower-quality preferences, or narrower semantic coverage. Update cost is there-

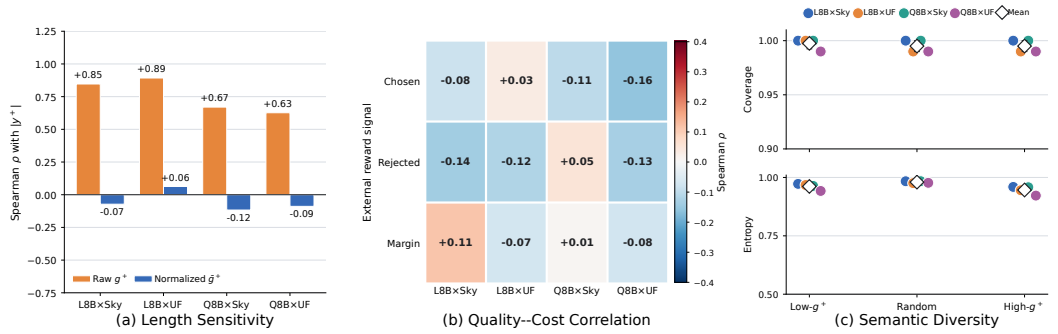

Figure 4: Chosen-side update cost and data attributes: (a) length sensitivity, (b) quality–cost correlation, and (c) semantic diversity.

fore not tightly coupled with these data attributes, leaving room to reduce it without systematically weakening preference supervision.

5 Preference Data Selection with Low-g^+

The KL-regularized DPO objective balances preference gain against policy-deviation cost, so offline data selection should account for both supervision and stability. The analyses above show that chosen-side updates more directly characterize generation-relevant local perturbations, while $\bar{g}^+$ is only weakly coupled with preference quality and Low-g^+ retains diversity comparable to Random. We therefore propose Low-g^+, which prioritizes preference pairs with lower chosen-side update cost without systematically weakening supervision.

Given a candidate dataset $\mathcal{D}$ and a data budget K, Low-g^+ ranks complete preference pairs by the length-normalized chosen-gradient score $\bar{g}^+$ defined above and selects the lowest-scoring subset:

$$\mathcal{D}_{\text{sel}} = \text{BottomK}_{(x,y^+,y^-)\in\mathcal{D}}\, \bar{g}^+(x, y^+). \tag{7}$$

The chosen-side gradient is used only for ranking; selected pairs retain (x, y^+, y^-) and use standard DPO. We estimate $\bar{g}^+$ with a temporary LoRA adapter (Hu et al., 2022), discard it after ranking, and run full-parameter DPO from the original SFT checkpoint; details are in Appendix E.

The stability motivation for Low-g^+ follows from how $\bar{g}^+$ controls the size of a local update. Let $\delta\theta^+$ denote the chosen-side component of a one-step gradient update with learning rate η. Because the DPO sample weight satisfies $0 < w_\theta \le \beta$ and $g^+ = \sqrt{|y^+|}\bar{g}^+$, if $|y^+| \le L_{\max}$, then

$$\|\delta\theta^+\|_2 \le \eta\beta g^+ \le \eta\beta\sqrt{L_{\max}}\,\bar{g}^+. \tag{8}$$

Thus, a smaller $\bar{g}^+$ directly yields a tighter upper bound on the chosen-side parameter update.

We next connect the parameter update to the generation distribution. Let $\mathcal{T}_\theta(x)$ denote a fixed high-probability generation region under the pre-update policy. Under local smoothness, the log-probability change of any response in this region satisfies

$$\sup_{y\in\mathcal{T}_\theta(x)} |\log \pi_{\theta+\delta\theta^+}(y \mid x) - \log \pi_\theta(y \mid x)| \le \underbrace{C_1\bar{g}^+ + C_2(\bar{g}^+)^2}_{\epsilon(\bar{g}^+)}, \tag{9}$$

where $C_1, C_2 \ge 0$ depend on the training configuration, response-length bound, and local properties of the current policy. Therefore, reducing $\bar{g}^+$ tightens not only the parameter-update bound but also the distribution-perturbation bound within the high-probability region.

Let $\mathcal{G}_\theta(x) \subseteq \mathcal{T}_\theta(x)$ denote the high-quality subset of this region, and let $\widetilde{\pi}_\theta$ denote the policy restricted and renormalized over $\mathcal{T}_\theta(x)$. The post-update probability mass of this high-quality subset satisfies

$$\widetilde{\pi}_{\theta+\delta\theta^+}\big(\mathcal{G}_\theta(x) \mid x\big) \ge e^{-2\epsilon(\bar{g}^+)}\widetilde{\pi}_\theta\big(\mathcal{G}_\theta(x) \mid x\big). \tag{10}$$

Provided that effective preference supervision is preserved, reducing $\bar{g}^+$ increases the factor $e^{-2\epsilon(\bar{g}^+)}$ and therefore raises the retained-mass lower bound for high-quality responses. Low-g^+

thus does not simply seek smaller model updates; it selects preference pairs with a more favorable local stability guarantee without systematically weakening preference quality or semantic coverage. This bound establishes the stability benefit of reducing local perturbation, while the experiments below test whether it translates into better alignment performance. Appendix D provides the full derivation and explicit forms of C_1 and C_2.

6 EXPERIMENTS

6.1 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. Detailed training settings are provided in Appendix E.

Preference Datasets. We use three publicly available preference datasets: Skywork (Liu et al., 2024) (70k pairs), a carefully curated mixture of open-source preference data; UltraFeedback (Cui et al., 2024) (57k pairs), covering diverse instruction types; and Nectar (Zhu et al., 2023) (123k pairs), with broad topic coverage.

Evaluation. We evaluate on three alignment benchmarks—AlpacaEval 2 (Dubois et al., 2024), MT-Bench (Zheng et al., 2023), and Arena-Hard (Li et al., 2024), with GPT-4 (OpenAI, 2023) as judge—and report zero-shot MMLU (Hendrycks et al., 2021) and MMLU-Pro (Wang et al., 2024) for general knowledge retention. AlpacaEval 2 and Arena-Hard report win rates against both SFT and official reference responses. Full evaluation and gradient-scoring details are in Appendix E.

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).

Data Scaling Protocol. To test whether the benefit is specific to the 30% main-experiment budget, we additionally use Llama-3.1-8B and vary the retained fraction from 10% to 100% in 10-point increments on Skywork and UltraFeedback. Low-g^+ and Random always receive the same data budget and are evaluated by AlpacaEval 2 win rate against the SFT reference. Appendix E further reports results at 20% and 50% selection ratios for three representative models across all three datasets.

6.2 MAIN RESULTS

Table 2 reports the main results across four models and three datasets at a 30% selection ratio. Low-g^+ ranks first or second in 55 of 60 metric entries (4 models $\times$ 3 datasets $\times$ 5 metrics), including 36 first-place results. The next-best methods are SelDPO (11 firsts, 10 seconds) and BeeS (7 firsts, 12 seconds).

The advantage holds across evaluation dimensions. On AlpacaEval vs SFT (AE_S), Low-g^+ ranks first or second in 11 of 12 model–dataset settings and exceeds Full, BeeS, and Random by 6.3, 4.0, and 7.1 points on average. It ranks first in 8 of 12 MT-Bench settings, with a mean score of 6.32 versus 6.07 for Full and 6.11 for BeeS. On Arena-Hard vs SFT (AH_S), it ranks first in 7 of 12 settings and exceeds Full by 1.8 points on average. The gains therefore persist across model families, data sources, and alignment benchmarks.

The baselines are more sensitive to the model–dataset combination. SelDPO is competitive on UltraFeedback but falls below Random on several Skywork settings, while Difficulty underperforms Random in most comparisons. BeeS remains competitive, but is less consistent than Low-g^+ across the five metrics. These variations indicate that the effectiveness of reward-margin, loss-based, and difficulty-based criteria depends on the model and offline-data distribution, reinforcing the need for a policy-dependent selection signal.

Table 2: Main results across four models and three preference datasets. AE_S/AE_R: AlpacaEval 2 win rate (%) vs SFT / official reference; MT: MT-Bench (1–10); AH_S/AH_R: Arena-Hard win rate (%) vs SFT / official reference. All selection methods use 30% of the data. **Bold** and underlined values denote the best and second-best results, respectively, based on unrounded scores.

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**	**7.70**	**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^+	**80.7**	**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^+	**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^+	**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

6.3 Analysis

Selection-Signal and Score-Range Analysis. Table 3 asks whether the chosen-gradient score is a useful ranking signal rather than merely a binary selector. We compare equal-budget subsets from low, middle, and high score ranges; Low-γ (full-gradient norm), High-cos (top cosine similarity between the chosen-side and rejected-side gradients), and Low-g^- (rejected-gradient norm) serve as alternative signals. R_{30}/3E is the standard 30% Random control trained for three epochs, whereas R_{90}/1E uses 90% of the data for one epoch. The latter approximately matches the number of sample presentations while increasing unique-data coverage.

The score-range trend is monotonic in all eight settings: the low-$\bar{g}^+$ subset is best, followed by the middle and then high-score subset. On average, selecting the low-score range improves AlpacaEval 2 win rate by 7.0 points over the middle range and 16.6 points over the high range. It also exceeds R_{30}/3E in every setting, showing that the trend is not explained by random subsampling. More importantly, Low-g^+ beats the equal-presentation R_{90}/1E control in all eight settings by 6.7 points on average. The gain therefore cannot be reduced to seeing more unique pairs: the ordering induced by the chosen-gradient score matters.

Among the alternative signals, Low-g^+ outperforms every selector in all eight settings. Its average gain is 5.0 points over Low-γ, 6.6 over High-cos, and 11.9 over Low-g^-. The advantage over Low-

g^- ranges from 2.6 to 25.1 points, supporting the chosen-side gradient as a more reliable stability-oriented ranking signal than the rejected-side gradient. The Low-γ comparison further shows that incorporating both sides into a full-gradient score does not improve ranking in these offline settings. Every selected subset retains complete preference pairs and uses the standard DPO objective, so these differences isolate data selection rather than a change in training objective. Appendix E finds no consistent DPO-loss ordering among the Low-, Middle-, and High-$\bar{g}^+$ subsets across datasets, indicating no strong systematic association between the chosen-gradient score and training loss. With matched sample presentations, Low→Low→Low outperforms both Low→Middle→High and High→Middle→Low in all eight settings, showing that the benefit persists beyond early training.

Table 3: AlpacaEval 2 win rate (%) against SFT for chosen-score ranges, alternative gradient signals, and random controls.

	Chosen-Score Range ($\bar{g}^+$)				**Alternative Signals and Random Control**			
Setting	**Low**	**Middle**	**High**	**R_{30}/3E**	**Low-γ**	**High-cos**	**Low-g^-**	**R_{90}/1E**
L3B×Sky	78.0	75.2	58.3	66.4	70.8	66.9	59.0	67.3
L3B×UF	85.7	81.1	73.2	79.7	80.0	76.9	76.6	81.1
L3B×Nec	83.5	79.3	77.4	80.4	81.1	80.6	79.5	79.3
L8B×Sky	80.7	70.8	58.0	65.8	80.2	73.6	62.7	72.8
L8B×UF	87.8	75.5	65.2	78.8	86.6	81.1	83.3	77.3
L8B×Nec	84.7	74.8	68.4	81.0	82.8	81.0	82.1	82.3
Q4B×Sky	68.9	64.6	60.1	67.0	62.8	66.5	56.0	64.1
Q4B×UF	88.2	80.4	64.2	77.3	73.4	78.1	63.1	79.6

Data Scaling. Figure 1 shows that the improvement is not an artifact of one 30% operating point: Low-g^+ exceeds Random at every non-full fraction on both datasets. At the 30% budget used in the main experiments, the gain is 14.9 points on Skywork (80.7 vs. 65.8) and 9.0 points on UltraFeedback (87.8 vs. 78.8). Low-g^+ reaches its best observed score with only 30% of Skywork and 40% of UltraFeedback, whereas Random remains below it until both methods use the full corpus. This low- and mid-budget advantage shows that the chosen-gradient score improves the data-efficiency frontier rather than merely identifying a favorable fixed-ratio subset.

Compute and General Capabilities. Appendix E also evaluates computational cost and general knowledge retention. On Llama-3.2-3B with Skywork, Low-g^+ uses approximately 47% of the FLOPs of full-data DPO when selection and training are counted together: more than Random (30%) because it requires one chosen-side scoring pass, but less than SelDPO (70%). Across MMLU (Hendrycks et al., 2021) and MMLU-Pro (Wang et al., 2024), its scores are generally close to the Full baseline, suggesting that the alignment gains are not accompanied by a substantial loss of broad general knowledge.

7 Conclusion

We present a chosen-response-centered approach to DPO data selection. Although the pairwise objective is symmetric, chosen-side updates directly reinforce high-probability generation patterns, whereas rejected-side effects depend on whether undesirable patterns remain likely. This motivates Low-g^+, which selects pairs with smaller length-normalized chosen-gradient scores and lower local update cost. Across four models and three datasets, Low-g^+ performs strongly with 30% of the data relative to full-data training and baselines. Controlled generation and gradient ablations support this mechanism; data scaling shows benefits across ratios.

Limitations

Low-g^+ depends on scores computed from the current policy, so substantial changes to the base checkpoint require rescoring. Although temporary LoRA-based scoring substantially reduces the cost, it still adds forward-backward passes beyond Random selection. Our experiments cover 3B–8B models and three offline preference datasets; validation on substantially larger models, additional preference domains, and iterative settings remains future work.

REFERENCES

Afra Amini, Tim Viber, and Ryan Cotterell. Direct preference optimization with an offset, 2024. URL https://arxiv.org/abs/2402.10571.

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

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

Angelica Chen, Sadhika Malladi, Lily H. Zhang, Xinyi Chen, and Qiuyi Zhang. Preference learning algorithms do not learn preference rankings, 2024a. URL https://arxiv.org/abs/2405.19534.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024b. URL https://arxiv.org/abs/2402.03216.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9722–9744. PMLR, 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Active preference optimization for sample efficient RLHF, 2024. URL https://arxiv.org/abs/2402.10500.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. Less is more: Improving LLM alignment via preference data selection, 2025. URL https://arxiv.org/abs/2502.14560.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023. URL https://arxiv.org/abs/2305.14233.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators, 2024. URL https://arxiv.org/abs/2404.04475.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization, 2024. URL https://arxiv.org/abs/2402.01306.

Duanyu Feng, Bowen Qin, Chen Huang, Zheng Zhang, and Wenqiang Lei. Towards analyzing and understanding the limitations of DPO: A theoretical perspective, 2024. URL https://arxiv.org/abs/2404.04626.

Chengqian Gao, Haonan Li, Liu Liu, Zeke Xie, Peilin Zhao, and Zhiqiang Xu. Principled data selection for alignment: The hidden risks of difficult examples. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18386–18409. PMLR, 2025. URL https://proceedings.mlr.press/v267/gao25f.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, and Scott Niekum. Contrastive preference learning: Learning from human feedback without RL, 2023. URL https://arxiv.org/abs/2310.13639.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model, 2024. URL https://arxiv.org/abs/2403.07691.

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

Kexin Huang, Junkang Wu, Ziqian Chen, Xue Wang, Jinyang Gao, Bolin Ding, Jiancan Wu, Xiangnan He, and Xiang Wang. Larger or smaller reward margins to select preferences for LLM alignment? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25922–25946. PMLR, 2025. URL https://proceedings.mlr.press/v267/huang25al.html.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-Hard and BenchBuilder pipeline, 2024. URL https://arxiv.org/abs/2406.11939.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, et al. Skywork-reward: Bag of tricks for reward modeling in LLMs, 2024. URL https://arxiv.org/abs/2410.18451.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-reward-v2: Scaling preference data curation via human–ai synergy, 2025. URL https://arxiv.org/abs/2507.01352.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning, 2023. URL https://arxiv.org/abs/2312.15685.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward, 2024. URL https://arxiv.org/abs/2405.14734.

OpenAI. GPT-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774.

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

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with DPO-positive, 2024. URL https://arxiv.org/abs/2402.13228.

Yu Pan, Zhongze Cai, Guanting Chen, Huaiyang Zhong, and Chonghuan Wang. What matters in data for DPO?, 2025. URL https://arxiv.org/abs/2508.18312.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization, 2024. URL https://arxiv.org/abs/2403.19159.

Xuan Qi, Rongwu Xu, and Zhijing Jin. Difficulty-based preference data selection by DPO implicit reward gap, 2025. URL https://arxiv.org/abs/2508.04149.

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

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16. IEEE, 2020. doi: 10.1109/SC41405.2020.00024. URL https://doi.org/10.1109/SC41405.2020.00024.

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

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47441–47474. PMLR, 2024. URL https://proceedings.mlr.press/v235/tajwar24a.html.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, and Will Dabney. Understanding the performance gap between online and offline alignment algorithms, 2024. URL https://arxiv.org/abs/2405.08448.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of LM alignment, 2023. URL https://arxiv.org/abs/2310.16944.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL https://arxiv.org/abs/2406.01574.

Yunan Wang, Jijie Li, Bo-Wen Zhang, Liangdong Wang, and Guang Liu. InCo-DPO: Balancing distribution shift and data quality for enhanced preference optimization, 2025. URL https://arxiv.org/abs/2503.15880.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning, 2024. URL https://arxiv.org/abs/2402.04333.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54715–54754. PMLR, 2024. URL https://proceedings.mlr.press/v235/xiong24a.html.

Yuzi Yan, Yibo Miao, Jialian Li, Yipin Zhang, and Jian Xie. 3D-properties: Identifying challenges in DPO and charting a path forward, 2024. URL https://arxiv.org/abs/2406.07327.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E. Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR, 2024. URL https://proceedings.mlr.press/v235/yuan24d.html.

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

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

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. Starling-7b: Improving LLM helpfulness & harmlessness with RLAIF. Nectar dataset card, November 2023. URL https://huggingface.co/datasets/berkeley-nest/Nectar.

A THE USE OF LARGE LANGUAGE MODELS (LLMS)

In the preparation of this work, we used LLMs as auxiliary tools in a limited capacity. Specifically, LLMs assisted in drafting portions of the code and in refining the wording of certain sentences for clarity and readability. All technical content, including the design of algorithms, experimental methodology, analysis, and interpretations, was independently developed by the authors. The use of LLMs was confined to language refinement and coding suggestions, and did not influence the scientific contributions or results reported in this paper.

B IDEAL DPO PROBABILITY-SHIFT BOUND

Fix a prompt x and let

$$\mathcal{S}_0(x) = \{y : \pi_0(y \mid x) > 0\} \tag{11}$$

be the support of the reference policy. We optimize over distributions supported on $\mathcal{S}_0(x)$ and assume $\beta > 0$. Consider the KL-regularized reward-maximization problem

$$\pi_r = \arg\max_{\pi} \left\{ \mathbb{E}_{y\sim\pi(\cdot|x)}[r(x,y)] - \beta D_{\mathrm{KL}}\big(\pi(\cdot \mid x) \,\|\, \pi_0(\cdot \mid x)\big) \right\}. \tag{12}$$

For a discrete response space, introduce a Lagrange multiplier λ for the normalization constraint. The Lagrangian is

$$\begin{aligned}\mathcal{L}(\pi,\lambda) = &\sum_{y\in\mathcal{S}_0(x)} \pi(y \mid x) r(x,y) - \beta \sum_{y\in\mathcal{S}_0(x)} \pi(y \mid x) \log \frac{\pi(y \mid x)}{\pi_0(y \mid x)} \\ &+ \lambda \left(\sum_{y\in\mathcal{S}_0(x)} \pi(y \mid x) - 1 \right).\end{aligned} \tag{13}$$

At an interior optimum, stationarity with respect to each $\pi(y \mid x)$ gives

$$r(x,y) - \beta \left(\log \frac{\pi(y \mid x)}{\pi_0(y \mid x)} + 1 \right) + \lambda = 0. \tag{14}$$

Rearranging,

$$\pi(y \mid x) = \pi_0(y \mid x) e^{r(x,y)/\beta} e^{\lambda/\beta - 1}. \tag{15}$$

The final factor is constant in y. Enforcing normalization therefore yields the non-parametric optimum

$$\pi_r(y \mid x) = \frac{\pi_0(y \mid x) e^{r(x,y)/\beta}}{Z_r(x)}, Z_r(x) = \mathbb{E}_{y'\sim\pi_0(\cdot|x)} \left[e^{r(x,y')/\beta} \right]. \tag{16}$$

Because the negative KL term is strictly concave on the interior of the probability simplex, this stationary point is the unique optimum on $\mathcal{S}_0(x)$. The same derivation extends to a general response space by replacing sums with integrals and assuming the partition function is finite.

Now let

$$m_x = \inf_{y\in\mathcal{S}_0(x)} r(x,y), \qquad M_x = \sup_{y\in\mathcal{S}_0(x)} r(x,y), \qquad M_x - m_x \le R. \tag{17}$$

Exponentiating the reward bounds gives

$$e^{m_x/\beta} \le e^{r(x,y)/\beta} \le e^{M_x/\beta}. \tag{18}$$

Taking the expectation under $\pi_0(\cdot \mid x)$ and using that its total mass is one yields

$$e^{m_x/\beta} \le Z_r(x) \le e^{M_x/\beta}. \tag{19}$$

For every $y \in \mathcal{S}_0(x)$, the lower ratio bound follows by combining the smallest possible numerator with the largest possible normalizer:

$$\frac{\pi_r(y \mid x)}{\pi_0(y \mid x)} = \frac{e^{r(x,y)/\beta}}{Z_r(x)} \ge \frac{e^{m_x/\beta}}{e^{M_x/\beta}} = e^{-(M_x - m_x)/\beta} \ge e^{-R/\beta}. \tag{20}$$

Similarly,

$$\frac{\pi_r(y \mid x)}{\pi_0(y \mid x)} \leq \frac{e^{M_x/\beta}}{e^{m_x/\beta}} = e^{(M_x - m_x)/\beta} \leq e^{R/\beta}. \tag{21}$$

Combining the two sides gives the pointwise probability-ratio bound:

$$e^{-R/\beta} \leq \frac{\pi_r(y \mid x)}{\pi_0(y \mid x)} = \frac{e^{r(x,y)/\beta}}{Z_r(x)} \leq e^{R/\beta}. \tag{22}$$

Multiplying the pointwise inequality by $\pi_0(y \mid x)$ and summing over any measurable response set A gives the corresponding setwise bound, whose lower side is used in the main paper:

$$e^{-R/\beta}\pi_0(A \mid x) \leq \pi_r(A \mid x) \leq e^{R/\beta}\pi_0(A \mid x). \tag{23}$$

If $\pi_0(A \mid x) = 0$, then $\pi_r(A \mid x) = 0$, so the optimum remains absolutely continuous with respect to the reference policy. The bound depends only on the reward range $M_x - m_x$ and is invariant to adding a prompt-dependent constant to every reward. A smaller ratio R/β restricts how strongly the ideal optimum can reweight the reference distribution.

This derivation concerns the ideal non-parametric optimum of the KL-regularized objective. It motivates a stability-oriented view of DPO, but it does not imply that every finite-step parametric update satisfies the same probability-ratio bound. The local parameter and generation-distribution bounds relevant to Low-g^+ are derived separately in Appendix D.

C Controlled One-Step Update Protocol

Models and data construction. We conduct the controlled analysis on Llama-3.2-3B, Llama-3.1-8B, and Qwen3-4B across Skywork, UltraFeedback, and Nectar, giving nine model–dataset settings. For each setting, we shuffle the preference data with seed 42 and retain 500 non-empty pairs whose chosen and rejected responses have similar character lengths:

$$\frac{\max(|y^+|_{\text{char}}, |y^-|_{\text{char}})}{\min(|y^+|_{\text{char}}, |y^-|_{\text{char}})} \leq 1.1. \tag{24}$$

This filter reduces response-length confounding when comparing the two gradient directions.

One-step update protocol. For each pair, we first greedily decode a baseline response from the SFT policy with at most 2048 new tokens. Starting from the same saved model parameters, we then compute the gradient of the summed response-token log-likelihood and apply either one full-parameter ascent step on y^+ or one full-parameter descent step on y^-. Both directions use learning rate $\eta = 10^{-5}$; the original parameters are restored before the other direction is evaluated. Inputs are truncated to 2048 tokens and forward/backward computation uses bf16. Thus, the two directions differ only in the response and sign of the controlled update.

Generation metrics. Let $\hat{y}$ be the baseline greedy output, and let $\hat{y}^+$ and $\hat{y}^-$ be the outputs after the chosen-side and rejected-side updates. We report

$$\Delta\mathrm{RL}^+ = \mathrm{RL}(\hat{y}^+, y^+) - \mathrm{RL}(\hat{y}, y^+), \tag{25}$$

and

$$\Delta\mathrm{RL}^- = \mathrm{RL}(\hat{y}^-, y^-) - \mathrm{RL}(\hat{y}, y^-), \tag{26}$$

where RL is token-level ROUGE-L (Lin, 2004) computed on at most the first 512 tokens. Positive $\Delta\mathrm{RL}^+$ indicates movement toward the chosen response, whereas negative $\Delta\mathrm{RL}^-$ indicates suppression of the rejected response. The first divergence position is the earliest token at which an updated output differs from the baseline; if no difference occurs before either sequence ends, we use the shorter sequence length. We also compute response-token-averaged log likelihoods under the original policy to compare the probability status of y^+ and y^-.

D Full Stability-Bound Derivation

We now connect the chosen-gradient score to the retained probability mass in a fixed, pre-update generation region. For a response set $\mathcal{T}_\theta(x)$ with $\pi_\theta(\mathcal{T}_\theta(x) \mid x) > 0$, define the policy restricted and renormalized to this region by

$$\widetilde{\pi}_\theta(y \mid x) = \frac{\pi_\theta(y \mid x)\mathbb{I}[y \in \mathcal{T}_\theta(x)]}{\sum_{y' \in \mathcal{T}_\theta(x)} \pi_\theta(y' \mid x)}. \tag{27}$$

The same pre-update region $\mathcal{T}_\theta(x)$ is used to normalize both the pre- and post-update policies. This fixed-region comparison avoids conflating probability perturbation with a change in region membership.

Proposition 1. *Let $\mathcal{G}_\theta(x) \subseteq \mathcal{T}_\theta(x)$ be any subset, let $\epsilon \geq 0$, and suppose that for every $y \in \mathcal{T}_\theta(x)$ with positive pre-update probability,*

$$|\log \pi_{\theta'}(y|x) - \log \pi_\theta(y|x)| \leq \epsilon, \tag{28}$$

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

$$\widetilde{\pi}_{\theta'}(\mathcal{G}_\theta(x) \mid x) \geq \exp(-2\epsilon)\, \widetilde{\pi}_\theta(\mathcal{G}_\theta(x) \mid x). \tag{29}$$

Proof. Subtracting the two log probabilities and exponentiating the assumed inequality gives, for every $y \in \mathcal{T}_\theta(x)$,

$$e^{-\epsilon} \leq \frac{\pi_{\theta'}(y|x)}{\pi_\theta(y|x)} \leq e^{\epsilon}. \tag{30}$$

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

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

$$\sum_{y \in \mathcal{G}_\theta(x)} \pi_{\theta'}(y|x) \geq e^{-\epsilon} \sum_{y \in \mathcal{G}_\theta(x)} \pi_\theta(y|x). \tag{31}$$

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

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

$$\sum_{y \in \mathcal{T}_\theta(x)} \pi_{\theta'}(y|x) \leq e^{\epsilon} \sum_{y \in \mathcal{T}_\theta(x)} \pi_\theta(y|x). \tag{32}$$

Step 3: Combining via normalization.

The normalized mass under θ' is

$$\widetilde{\pi}_{\theta'}(\mathcal{G}_\theta(x) \mid x) = \frac{\sum_{y \in \mathcal{G}_\theta(x)} \pi_{\theta'}(y|x)}{\sum_{y \in \mathcal{T}_\theta(x)} \pi_{\theta'}(y|x)}. \tag{33}$$

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

$$\widetilde{\pi}_{\theta'}(\mathcal{G}_\theta(x) \mid x) \geq \frac{e^{-\epsilon} \sum_{y \in \mathcal{G}_\theta(x)} \pi_\theta(y|x)}{e^{\epsilon} \sum_{y \in \mathcal{T}_\theta(x)} \pi_\theta(y|x)} = e^{-2\epsilon}\, \widetilde{\pi}_\theta(\mathcal{G}_\theta(x) \mid x). \tag{34}$$

This completes the proof. □

Chosen-side parameter-update bound. For one preference pair, define the DPO logit

$$z_\theta = \beta \left[\log \frac{\pi_\theta(y^+ \mid x)}{\pi_0(y^+ \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_0(y^- \mid x)} \right]. \tag{35}$$

For the loss $\ell_{\mathrm{DPO}} = -\log \sigma(z_\theta)$, the scalar multiplying the two score-function components in the descent direction is

$$w_\theta = \beta \sigma(-z_\theta). \tag{36}$$

Since $0 < \sigma(-z_\theta) < 1$, we have $0 < w_\theta < \beta$. Isolating the likelihood-increasing chosen-side component of a gradient step with learning rate η gives

$$\delta\theta^+ = \eta w_\theta \nabla_\theta \log \pi_\theta(y^+ \mid x). \tag{37}$$

Consequently,

$$\|\delta\theta^+\|_2 = \eta w_\theta \left\|\nabla_\theta \log \pi_\theta(y^+ \mid x)\right\|_2 = \eta w_\theta g^+ \le \eta\beta g^+. \tag{38}$$

This is a bound on the chosen-side component used to motivate the ranking score; the actual DPO update still contains both chosen-side and rejected-side components.

Perturbation bound derivation. We derive the uniform perturbation bound used in Proposition 1. For each response $y \in \mathcal{T}_\theta(x)$, let $f_y(\theta) = \log \pi_\theta(y \mid x)$ and assume that there is a finite constant $L_\mathcal{T}$ such that ∇f_y is $L_\mathcal{T}$-Lipschitz along the line segment between θ and $\theta + \delta\theta^+$, uniformly over this region. The standard smoothness inequality gives

$$\left| f_y(\theta + \delta\theta^+) - f_y(\theta) - \left\langle \nabla f_y(\theta), \delta\theta^+ \right\rangle \right| \le \frac{L_\mathcal{T}}{2} \|\delta\theta^+\|_2^2. \tag{39}$$

Define

$$G_\mathcal{T} = \sup_{y \in \mathcal{T}_\theta(x)} \|\nabla_\theta \log \pi_\theta(y \mid x)\|_2. \tag{40}$$

We assume that $G_\mathcal{T} < \infty$ in this local region. By the triangle and Cauchy–Schwarz inequalities, the worst-case change over the fixed high-probability region satisfies

$$\sup_{y \in \mathcal{T}_\theta(x)} |\log \pi_{\theta+\delta\theta^+}(y \mid x) - \log \pi_\theta(y \mid x)| \le G_\mathcal{T} \|\delta\theta^+\|_2 + \frac{L_\mathcal{T}}{2} \|\delta\theta^+\|_2^2 \tag{41}$$

Because $g^+ = \sqrt{|y^+|}\, \bar{g}^+$, a response-length bound $|y^+| \le L_{\max}$ implies

$$\|\delta\theta^+\|_2 \le \eta\beta g^+ \le \eta\beta \sqrt{L_{\max}}\, \bar{g}^+. \tag{42}$$

Substitution yields

$$\begin{aligned} &\sup_{y \in \mathcal{T}_\theta(x)} |\log \pi_{\theta+\delta\theta^+}(y \mid x) - \log \pi_\theta(y \mid x)| \\ &\quad \le G_\mathcal{T} \eta\beta \sqrt{L_{\max}}\, \bar{g}^+ + \frac{L_\mathcal{T}}{2} \eta^2 \beta^2 L_{\max} (\bar{g}^+)^2 =: \epsilon(\bar{g}^+). \end{aligned} \tag{43}$$

where

$$C_1 = G_\mathcal{T} \eta\beta \sqrt{L_{\max}}, \qquad C_2 = \frac{L_\mathcal{T}}{2} \eta^2 \beta^2 L_{\max}. \tag{44}$$

Thus,

$$\epsilon(\bar{g}^+) = C_1 \bar{g}^+ + C_2 (\bar{g}^+)^2. \tag{45}$$

Both constants are non-negative for a fixed prompt, local region, checkpoint, and training configuration. Hence

$$\frac{d\epsilon(\bar{g}^+)}{d\bar{g}^+} = C_1 + 2C_2 \bar{g}^+ \ge 0 \qquad (\bar{g}^+ \ge 0), \tag{46}$$

so a smaller score gives a tighter perturbation upper bound. Applying Proposition 1 with

$$\epsilon(\bar{g}^+) = C_1 \bar{g}^+ + C_2 (\bar{g}^+)^2 \tag{47}$$

yields the explicit retained-mass bound

$$\widetilde{\pi}_{\theta+\delta\theta^+}\big(\mathcal{G}_\theta(x) \mid x\big) \ge \exp\big(-2C_1 \bar{g}^+ - 2C_2 (\bar{g}^+)^2\big)\, \widetilde{\pi}_\theta\big(\mathcal{G}_\theta(x) \mid x\big). \tag{48}$$

Thus, lowering $\bar{g}^+$ simultaneously tightens the chosen-side parameter-update bound, tightens the worst-case log-probability perturbation bound on the fixed generation region, and raises the resulting lower bound on retained normalized mass.

Scope of the result. The derivation is local: it analyzes one chosen-side component around the current checkpoint, assumes smoothness on the line segment traversed by that update, and keeps the pre-update region $\mathcal{T}_\theta(x)$ fixed. It does not claim that the rejected-side component is always negligible, that the full multi-step DPO trajectory obeys the same constants, or that a smaller gradient alone guarantees higher response quality. The empirical quality–cost and diversity analyses therefore provide complementary evidence that selecting lower-cost pairs does not systematically weaken preference supervision.

E Reproducibility and Additional Results

Code availability. Our code is available at the anonymous repository `https://anonymous.4open.science/status/Low-G-AF31`.

E.1 Training Configuration

All DPO experiments use full-parameter training (no LoRA) with $\beta = 0.1$. We train for three epochs using DeepSpeed ZeRO-3 (Rajbhandari et al., 2020) on eight NVIDIA A800 80 GB GPUs. The effective batch size is 128 (2 per device $\times$ 8 GPUs $\times$ 8 gradient accumulation steps). We use a cosine learning-rate schedule with a warmup ratio of 0.1. The learning rate is 2×10^{-6} for 3B and 4B models, and 5×10^{-7} for 8B models. The maximum sequence length is 2048 tokens, and training uses bf16 mixed precision. All baseline methods share the same training configuration to ensure fair comparison; only the data subset differs.

E.2 Gradient Scoring and Selection

Computing per-sample chosen-gradient norms over the full model is expensive. We adopt a lightweight proxy by enabling LoRA (Hu et al., 2022) on all linear layers with rank $r = 8$ and scaling parameter $\alpha = 16$. We first perform a short SFT warm-up on the candidate data and then compute the length-normalized chosen-gradient score $\bar{g}^+$ through the warmed-up adapter only. Scoring is performed on an NVIDIA A800 GPU. This replaces full-model per-sample backpropagation with adapter-only backpropagation and makes scoring datasets with over 100,000 samples practical.

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

$$\bar{g}^+(x, y^+) = \frac{1}{\sqrt{|y^+|}} \left\| \nabla_{\theta_{\mathrm{LoRA}}} \log \pi_\theta(y^+|x) \right\|, \tag{49}$$

where $|y^+|$ is the number of tokens in the chosen response and θ_{LoRA} denotes the adapter parameters. The $\frac{1}{\sqrt{|y^+|}}$ normalization mitigates the systematic increase in gradient scale caused by 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. Smaller values indicate smaller chosen-side proxy updates; Appendix D shows how the corresponding full-parameter score controls the local perturbation bound.

The LoRA adapter is discarded after scoring. Actual DPO training is full-parameter from the original SFT checkpoint, so the temporary adapter updates are not carried into final training.

E.3 LoRA Proxy Validation

To validate the scoring approximation, we randomly sample 5,000 preference pairs from each model–dataset setting and compute both the LoRA-proxy and full-parameter length-normalized chosen-gradient scores on the same pairs.

Table 4: Spearman rank correlation ρ between the LoRA-proxy and full-parameter length-normalized chosen-gradient scores.

Model	Skywork	UltraFeedback	Nectar
Llama-3.2-3B	0.97	0.93	0.95
Llama-3.1-8B	0.94	0.91	0.87
Qwen3-4B	0.93	0.89	0.92

The correlations range from 0.87 to 0.97, with an average of 0.92 across the nine settings. This consistently high rank agreement shows that the lightweight LoRA score closely tracks the full-parameter ordering across model architectures and data sources, supporting its use as an efficient proxy for data selection.

E.4 External Reward and Semantic Diversity Diagnostics

Quality–cost analysis. The external reward model used only for diagnostic analysis is `Skywork/Skywork-Reward-V2-Llama-3.1-8B` (Liu et al., 2025). It is not used for Low-g^+ ranking, preference-pair construction, or DPO training. For every candidate pair, we compute

$$q^+ = R(x, y^+), \qquad q^- = R(x, y^-), \qquad m = R(x, y^+) - R(x, y^-), \tag{50}$$

and report the Spearman correlations of $\bar{g}^+$ with chosen quality q^+, rejected quality q^-, and preference margin m. The analysis covers the full candidate corpora for Llama-3.1-8B and Qwen3-8B on Skywork and UltraFeedback; the reward model is never used as a data-selection signal.

Semantic diversity analysis. We concatenate each instruction with its chosen response into a single text sequence and encode it with `BAAI/bge-m3` (Chen et al., 2024b). For each of these four model–dataset settings, we apply K-means with $K = 100$ to the resulting representations. We compare equally sized Low-g^+, Random, and High-g^+ subsets, each containing 30% of the corpus. If n_k denotes the number of selected samples assigned to cluster k, we report cluster coverage and normalized cluster entropy:

$$\text{Coverage} = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}[n_k > 0], \tag{51}$$

$$\text{Entropy} = -\frac{1}{\log K} \sum_{k:n_k>0} \frac{n_k}{\sum_j n_j} \log \frac{n_k}{\sum_j n_j}. \tag{52}$$

Higher values indicate broader cluster coverage and a more even distribution over semantic clusters. These diagnostics test whether low update cost is obtained by narrowing semantic coverage; they do not enter the selection objective.

E.5 Evaluation Details

AlpacaEval 2. We generate responses on 805 AlpacaEval 2 (Dubois et al., 2024) instructions using greedy decoding (temperature 0, max 2048 tokens). We compute pairwise win rates against both the corresponding SFT baseline outputs and the official reference responses provided by the benchmark, with GPT-4 (OpenAI, 2023) judging each comparison.

MT-Bench. We use the standard 80-question MT-Bench (Zheng et al., 2023) 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. We generate responses on 500 challenging Arena-Hard (Li et al., 2024) prompts and compute pairwise win rates against both the corresponding SFT baseline outputs and the official reference responses provided by the benchmark, with GPT-4 judging each comparison.

Each model–dataset–method configuration is run three times with different random seeds, and the reported value is the mean over the three runs. For every model–dataset setting, the Full-data baseline performs best at epoch 3; we therefore report epoch-3 results for all methods. MMLU (Hendrycks et al., 2021) and MMLU-Pro (Wang et al., 2024) are evaluated in a zero-shot setting with our custom evaluation script; no in-context demonstrations are provided, and the same prompts and scoring implementation are used for every method. These scores assess whether data selection affects general knowledge capabilities.

E.6 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 response length of $T = 512$ tokens and three training epochs. Since each preference pair contains one chosen and one rejected response, DPO training processes approximately $2T$ response tokens per pair. We use the standard approximation of $6P$ FLOPs per token for a forward-backward pass. Prompt-token computation is omitted to keep a consistent response-token-based approximation across methods, so the reported FLOPs should be interpreted as coarse relative estimates. For

Difficulty and BeeS, we approximate selection as one forward-scoring pass over both responses. The BeeS estimate excludes additional external-reward-model passes and should therefore be interpreted as a lower-bound proxy.

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

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%

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) and the simplified estimates for Difficulty (Qi et al., 2025) and BeeS (Deng et al., 2025) (41%), 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). Together with its strong alignment results in the main paper, this yields a favorable performance–efficiency trade-off.

Table 6: Approximately compute-matched AlpacaEval 2 win rates (%) against SFT with Llama-3.1-8B. Including selection and training, Low-g^+ at 30% uses roughly 47% of full-training FLOPs, while Random at 50% uses 50%.

Dataset	**Low-g^+ 30%**	**Random 50%**	**Gain**
Skywork	80.7	66.6	+14.1
UltraFeedback	87.8	80.1	+7.7

The approximately compute-matched comparison further accounts for the scoring overhead. Despite using slightly fewer estimated total FLOPs, Low-g^+ at 30% exceeds Random at 50% by 14.1 points on Skywork and 7.7 points on UltraFeedback. Thus, its data-efficiency advantage persists when selection cost is included rather than comparing only the DPO training budgets.

E.7 Training Loss Across Score Ranges

Table 7: Mean DPO training loss logged within each epoch for Llama-3.1-8B runs on the Low, Middle, and High $\bar{g}^+$ score ranges.

Dataset	**Range**	**E1**	**E2**	**E3**
Skywork	Low	0.411	0.121	0.043
	Middle	0.409	0.113	0.037
	High	0.384	0.117	0.047
UltraFeedback	Low	0.527	0.237	0.121
	Middle	0.583	0.301	0.173
	High	0.607	0.355	0.236

On Skywork, the three score ranges exhibit similar loss trajectories, and the High-$\bar{g}^+$ subset has the lowest average loss in the first epoch. On UltraFeedback, the Low-$\bar{g}^+$ subset consistently attains lower training loss. The relationship between $\bar{g}^+$ and DPO training loss is therefore dataset dependent, indicating that the chosen-gradient score is not simply a loss-based measure of example difficulty. Moreover, the loss continues to decrease for all three subsets throughout training.

E.8 Curriculum Learning Across Score Ranges

For curriculum training, each stage uses one 30% $\bar{g}^+$ score range. The optimizer state and global cosine learning-rate schedule are preserved across stage transitions without restarting warmup. Consequently, the three-stage curricula, static Low-g^+ training, and R_{90}/1E use the same total number of sample presentations.

Table 8: AlpacaEval 2 win rate (%) against SFT for static Low-g^+ selection, Low-to-High and High-to-Low score curricula, and the equal-presentation Random control. L, M, and H denote the Low, Middle, and High $\bar{g}^+$ 30% score ranges, respectively; each curriculum uses one range per stage.

Setting	**L→L→L**	**L→M→H**	**H→M→L**	**R_{90}/1E**
L3B×Sky	78.0	74.2	60.7	67.3
L3B×UF	85.7	80.7	74.3	81.1
L3B×Nec	83.5	77.5	75.3	79.3
L8B×Sky	80.7	71.3	60.5	72.8
L8B×UF	87.8	76.3	66.8	77.3
L8B×Nec	84.7	75.3	69.2	82.3
Q4B×Sky	68.9	63.3	58.3	64.1
Q4B×UF	88.2	81.7	67.2	79.6

Static Low-g^+ selection is best in all eight settings, exceeding the Low-to-High and High-to-Low curricula by 7.2 and 15.7 points on average, respectively. Low-to-High also outperforms High-to-Low in every setting by 8.5 points on average, showing that starting from lower-cost samples is preferable in these curriculum settings. However, Low-to-High remains comparable to the equal-presentation Random control (75.0 vs. 75.5 on average) and does not recover the advantage of repeatedly training on the Low-g^+ subset. These results indicate that the benefit of Low-g^+ is not confined to the beginning of training: replacing lower-cost samples with progressively higher-cost ranges weakens final alignment performance under the same number of sample presentations.

E.9 Results at Other Selection Ratios

Tables 9 and 10 report results at 20% and 50% selection ratios on three representative models, complementing the 30% results in the main text.

Table 9: Results at a 20% selection ratio. Notation follows the main-results table. **Bold** and underlined values denote the best and second-best results, respectively, based on unrounded scores.

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	53.7	5.73	5.81	53.5	11.2	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^+	**77.3**	**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

Table 9 (continued)

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
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 10: Results at a 50% selection ratio. Notation follows the main-results table. **Bold** and underlined values denote the best and second-best results, respectively, based on unrounded scores.

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^+	**79.5**	**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	66.6	8.07	6.05	55.4	15.3	80.1	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

E.10 General Knowledge Retention

To assess whether data selection affects general knowledge capabilities, we report MMLU and MMLU-Pro scores for all methods across the three selection ratios. As shown in Tables 11–13, Low-g^+ remains close to the Full baseline in most settings, suggesting no substantial loss of broad knowledge capabilities.

Table 11: 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^+	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 12: 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^+	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

Table 12 (continued)

Method	Skywork		UltraFeedback		Nectar	
	MMLU	MMLU-Pro	MMLU	MMLU-Pro	MMLU	MMLU-Pro
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 13: MMLU and MMLU-Pro scores (%) at 50% 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	57.2	28.3	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	58.3	30.6	57.7	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	65.3	34.9
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