LC-DPO: Length-Controlled Direct Preference Optimization via Dynamic Token Selection

1st Jinlei Dong
School of Software Engineering
Xi'an Jiaotong University
Xi'an, China
dongjinlei@stu.xjtu.edu.cn

2nd Bin Zhang*
School of Software Engineering
Xi'an Jiaotong University
Xi'an, China
bzhang82@xjtu.edu.cn

3rd Jialong Zhang
School of Software Engineering
Xi'an Jiaotong University
Xi'an, China
zhangjialong@stu.xjtu.edu.cn

***Abstract*—Direct Preference Optimization often exhibits a pronounced verbosity bias, favoring longer responses even when quality does not improve. We propose LC-DPO, a simple extension of DPO that mitigates this bias by reweighting token level contributions on the longer response based on a policy reference loss gap criterion. Intuitively, LC-DPO emphasizes tokens that are more discriminative for preference learning under this criterion, while softly downweighting tokens that contribute weaker preference signal and are more correlated with verbosity. We evaluate LC-DPO on UltraFeedback with multiple backbones, including Llama-3-8B, Qwen2.5-7B, Mistral-8B, and Qwen2.5-1.5B. Results on MT-Bench and AlpacaEval2 show that LC-DPO consistently outperforms standard DPO and representative length control baselines, while producing shorter responses on average. These findings suggest LC-DPO is an effective and practical approach for reducing verbosity bias in preference alignment.**

***Index Terms*—Direct Preference Optimization, Length Bias, Reinforcement Learning, Large Language Models**

I. Introduction

Large Language Models (LLMs) can generate human-like text, handle complex contexts, and achieve strong generalization [1]. Modern LLM training typically involves pre-training, supervised fine-tuning, and preference alignment [2]. Among alignment techniques, Proximal Policy Optimization (PPO) is widely used to better match models with human preferences and improve instruct-following, usability, and safety [3]–[5].

Despite its effectiveness, PPO-based RLHF is complex and unstable. It usually requires training a separate reward model [6], [7] and then optimizing the policy with reinforcement learning [8]. This pipeline is sensitive to RM quality, while accurately modeling nuanced human preferences is difficult and often leads to unstable training dynamics [6], [9].To simplify preference alignment, [10] proposed Direct Preference Optimization (DPO), which directly optimizes a pairwise logistic objective on offline preference data. DPO avoids explicit reward modeling and on-policy RL [7], and has become a popular alignment method due to simplicity and stability [11].

However, DPO suffers from a notable verbosity bias: models fine-tuned with DPO often generate overly long responses [4]. This can hurt instruction adherence and reasoning quality while increasing inference cost [12]. Although part of the issue may stem from data bias: annotators and automatic evaluators sometimes prefer longer answers that appear more comprehensive [13], recent work suggests a more fundamental cause: DPO itself is sensitive to response length [14]. Since DPO's implicit reward aggregates log-likelihood over the whole sequence, longer preferred responses can contribute more accumulated terms to the optimization signal, encouraging the model to treat length as a proxy for preference [15].

To mitigate verbosity in DPO, prior work has explored several directions. Some methods modify the DPO objective, e.g., adding length regularization (R-DPO) or a length-controlled boundary loss (LMPO) [13], [15]. Others redefine the implicit reward, such as length normalization (SimPO) [16] or additional length-related reward heads (ALBM) [17]. At the token level, truncation or sampling strategies (LD-DPO, SamPO) can reduce the influence of long responses but may discard useful information [14], [18]. Recent token-level extensions of DPO, including TIS-DPO [19], TI-DPO [20], and OTPO [21], show that refining token contributions can improve preference learning. However, their primary focus is not on the systematic verbosity bias induced by response-length imbalance, leaving length-aware preference optimization insufficiently explored.

In this paper, we propose Length-Controlled Direct Preference Optimization (LC-DPO), a simple and effective method for mitigating verbosity bias in DPO. LC-DPO dynamically reweights tokens based on the policy–reference *Token Loss Gap*, emphasizing preference relevant tokens while downweighting less informative verbose ones. Experiments on UltraFeedback across multiple LLMs and benchmarks show that LC-DPO consistently outperforms standard DPO and other baselines, while reducing average response length by 13%–20%(Fig. 1). Our code is available at Github: https://github.com/Paipaier/lc_dpo. Our main contributions are as follows:

1) We analyze token-level loss dynamics in DPO and show that length imbalance is closely related to verbosity bias.
2) We propose LC-DPO, a simple length-controlled variant of DPO that assigns dynamic token weights using policy–reference loss gaps.
3) Experiments on multiple backbones and benchmarks show that LC-DPO consistently improves the length quality trade-off over DPO and representative baselines.

* Corresponding author.

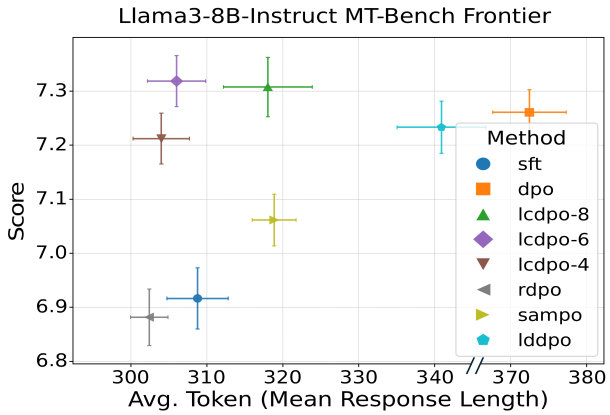

Fig. 1. MT-Bench performance of Llama 3-8B-Instruct under different training methods. LC-DPO variants (lcdpo-4/6/8) correspond to different values of the reweighting parameter α.

II. Method

In this section, we first provide a brief overview of the preliminaries. We then analyze the evolution of token-level loss during the DPO training process. Finally, we propose LC-DPO, which uses dynamic token selection to apply length regularization to the biased reward.

A. Preliminary Background of DPO

Given a dataset D composed of triplets of prompt x, a chosen y_w, and a rejected y_l, DPO directly optimizes the policy via an implicit Kullback-Leibler(KL) divergence, which increases the likelihood of y_w while decreasing the likelihood of y_l. The loss function of DPO can be expressed as follows:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma\left(\beta \triangle\right)\right] \tag{1}$$

$$\triangle = \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \tag{2}$$

where π_θ is the policy model, π_{ref} is the reference model, both of which are initialized by the model after Supervised Fine-Tuning (SFT). β is the hyperparameter that controls KL strength, and σ is the sigmoid activation function. Then we can express Eq. 2 at the token level as follows:

$$\Delta = \sum_{i=1}^{T_w} \log \frac{\pi_\theta(y_w^i|x)}{\pi_{ref}(y_w^i|x)} - \sum_{i=1}^{T_l} \log \frac{\pi_\theta(y_l^i|x)}{\pi_{ref}(y_l^i|x)} \tag{3}$$

The gradient at the token level loss can then be written as:

$$\nabla_\theta \mathcal{L}_{dpo}(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\beta\sigma(-\Delta)\mathcal{G}\right] \tag{4}$$

$$\mathcal{G} = \nabla_\theta \sum_{i=1}^{T_w} \log \pi(y_w^i|x) - \nabla_\theta \sum_{i=1}^{T_l} \log \pi(y_l^i|x) \tag{5}$$

where T denotes the response length. When y_w and y_l have comparable lengths, DPO can learn preference signals under relatively balanced token-level contributions. However, when y_w is substantially longer than y_l, the longer response may contribute more token-level terms to gradient, making the implicit preference reward more susceptible to length bias and encouraging longer outputs.

B. Not All Tokens Are Required in DPO

1) Motivation: length imbalance in preference pairs: Although DPO is optimized on paired preferences, its objective aggregates token-level terms over the full responses, making it sensitive to systematic length differences between chosen and rejected samples. To verify this premise on our training data, we measure response lengths on UltraFeedback using the Qwen2.5 tokenizer. Let T_w and T_l denote the token lengths of the chosen and rejected responses, and define $\Delta T = T_w - T_l$. On 61,135 training pairs, the chosen response is longer in 54.84% of cases (equal-length 1.78%), and ΔT is significantly shifted to the positive side (mean 37.3, median 12) with a heavy tail ($p_{80} = 303$, $p_{90} = 415$, $p_{99} = 649$).

Fig. 3(a) visualizes the magnitude of token-length differences, intuitively. The chosen-longer side exhibits both a larger median (120 vs 87) and a heavier long-length tail, indicating that length imbalance is not limited to marginal differences but is amplified by a subset of substantially longer responses. This non-negligible imbalance motivates explicitly calibrating token aggregation by the common length and reducing the contribution of length-driven tokens on the longer side.

2) Token loss dynamics: Inspired by RHO-1 [26], we introduce the concept of token-level negative log-likelihood (Token Loss). For each token x_i, the loss is defined as

$$\mathcal{L}(x_i) = -\log P_\theta(x_i \mid x_{<i}). \tag{6}$$

We re-examine how token losses evolve during DPO training by fine-tuning Llama-3-8B-Instruct model on UltraFeedback [4] and saving intermediate checkpoints. Fig. 2(a) shows that while most tokens remain consistently low-loss during training, a minority exhibit pronounced loss changes. Moreover, token losses fluctuate during training with substantially different magnitudes across tokens (Fig. 2(b)), suggesting that a subset of tokens contributes disproportionately to the optimization signal and dynamically evolving over training. This observation motivates a dynamic token selection or reweighting mechanism, which we formalize in the next section.

3) Loss-gap distribution on common vs. extra tokens: To further justify the use of loss-gap based token selection, we analyze how token-level loss gaps distribute across the longer response. For each preference pair, we consider only the longer response. We split tokens on the longer response into a common set (Top-Tp by loss-gap) and a extra set (the remaining tokens). Fig. 3 (b) plots the cumulative distributions of g_i (in log scale for better observation) for common and extra tokens on UltraFeedback. Extra tokens exhibit a clear left-shifted distribution, with substantially smaller gaps across the entire range. For instance, the median gap of extra tokens is less than half that of common tokens, and the difference persists at higher quantiles (e.g., p90). Statistical tests further confirm the separation (KS statistic $= 0.101$, $p < 10^{-16}$). These results indicate that tokens not in Top-Tp loss gap tend to carry weaker policy–reference disagreement, and are therefore less informative for preference learning. This empirical observation supports our design choice to downweight extra tokens on the longer response, while prioritizing tokens with larger loss gaps.

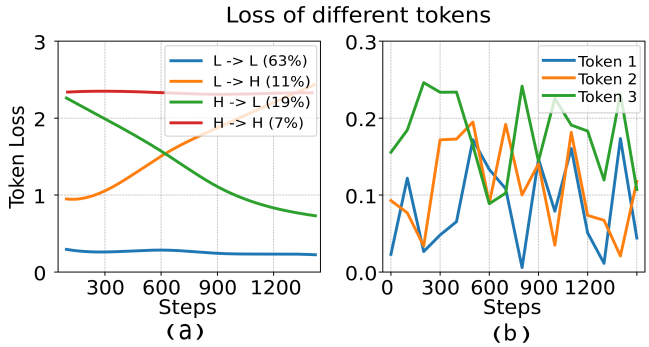

Fig. 2. Change in token loss during DPO. (a) presents the different types of tokens: L→L (63%), L→H (11%), H→L (19%), and H→H (7%), where H or L indicates high loss or low loss. (b) illustrates the change in several token loss during the training process.

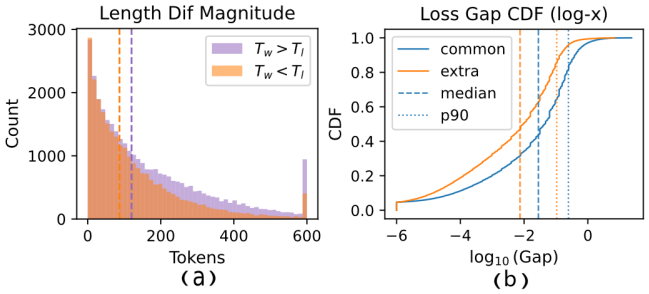

Fig. 3. (a) Magnitude of token-length differences between chosen and rejected responses in UltraFeedback. Dashed lines mark the median, respectively. (b) CDF of token-level loss gaps on the longer response (log-x).

C. Length-Controlled DPO

We now introduce our method: **L**ength-**C**ontrolled **D**irect **P**reference **O**ptimization via Dynamic Token Selection. Our approach mitigates length-related bias in DPO through dynamic token reweighting. For a given preference tuple (x, y_w, y_l). We define $T_p = \min(T_w, T_l)$ as the length of the shorter response, which we use as a length-matched quota when constructing token-level weights for the longer response.

We use a selection mechanism based on a Token Loss Gap to determine which tokens are more likely to carry preference-relevant signal and which are related to length-correlated signal. The algorithm's architecture is shown in Fig. 4. During the DPO training process, the token loss under the reference model can be expressed as:

$$\mathcal{L}_{ref}(y_i) = -\log P(y_i|x, y_{<i}) \tag{7}$$

Also, we can get the loss under the policy model:

$$\mathcal{L}_{\theta}(y_i) = -\log P(y_i|x, y_{<i}) \tag{8}$$

Then, we can get the Token Loss Gap between the current policy model and the reference model, which we define as g:

$$g(y_i) = |\mathcal{L}_{\theta}(y_i) - \mathcal{L}_{ref}(y_i)| \tag{9}$$

We sample the same number of tokens from the chosen and rejected responses. For the longer response, we select the Top-T_p tokens based on token loss gap g to represent the preference-relevant tokens, while the remaining token-level weight multiplying the original token loss term. At the same time, to prevent strong semantic discontinuities, we introduce a hyperparameter α to control the degree of length preference. The loss for LC-DPO is:

$$\mathcal{L}_{LC-DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log \sigma(\beta \triangle)\right] \tag{10}$$

$$\triangle = \left(\sum_{i=1}^{T_w} I_k(y_w^i) \log \frac{\pi_\theta(y_w^i|x)}{\pi_{\text{ref}}(y_w^i|x)}\right) - \left(\sum_{i=1}^{T_l} I_k(y_l^i) \log \frac{\pi_\theta(y_l^i|x)}{\pi_{\text{ref}}(y_l^i|x)}\right) \tag{11}$$

where $I_k(y_i) = \begin{cases} 1 & \text{if } y_i \in \text{Top-}T_p \text{ by g}(y_i) \\ \alpha & \text{otherwise} \end{cases}$, when $\alpha = 1$, the algorithm is identical to the original DPO. When $\alpha = 0$, the likelihood of the extra tokens becomes zero, so only the common length part (Top-Tp by loss gap) is considered. When DPO training is highly sensitive to response length or the preference data exhibits a pronounced length imbalance, a smaller α can be used to more strongly suppress length-correlated token contributions. Conversely, a larger α is preferable when aiming to preserve more preference-relevant signal and avoid overly aggressive downweighting. Here, T_p is not treated as the true informative length of a response, but as a conservative length-matched token budget for calibrating the longer response within each preference pair.

In terms of computation, LC-DPO does not introduce additional policy or reference forward passes beyond standard DPO. The extra operations are limited to token-wise loss subtraction, Top-Tp selection on the longer response, and mask construction, all of which are lightweight element-wise operations. Therefore, the added overhead is negligible relative to the full forward or backward cost of DPO training.

III. Experiment Setups

1) Models and Datasets: Our experiments are conducted on four widely adopted base models: Llama-3-8B, Qwen2.5-7B, Mistral-8B, and Qwen2.5-1.5B. We use the open-source UltraFeedback dataset, which provides 61k high-quality and diverse human preference pairs aimed at improving helpfulness and harmlessness.

2) Baselines and Evaluation Benchmarks: We compared LC-DPO with five other preference optimization methods: DPO, R-DPO, SimPO, SamPO, and LD-DPO. We evaluate our model on two popular open-ended benchmarks: MT-Bench and AlpacaEval2. MT-Bench includes over 80 questions in 8 categories to assess performance in various real-world scenarios. AlpacaEval2 focuses on instruction following and measures the win rate against GPT-4 with controlled response lengths. We also calculate the average response length to compare the effectiveness of different methods. In addition, we test LC-DPO on six conditional generation benchmarks: GSM8K [27], IFEval [28], PiQA [29], MMLU [30], COQA [31], and HellaSwag [32], which evaluate abilities ranging from arithmetic and commonsense reasoning to factual accuracy.

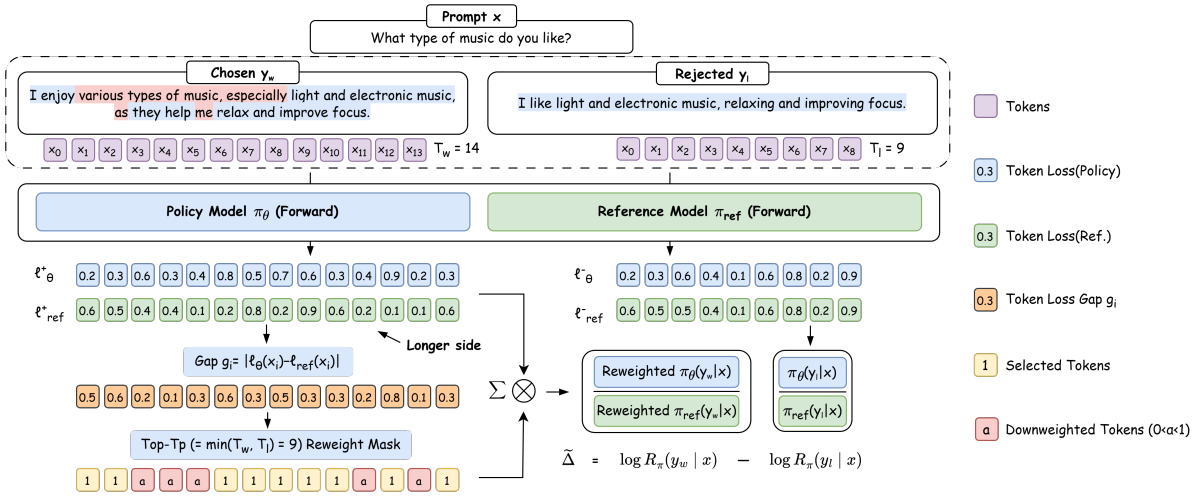

Fig. 4. Overview of LC-DPO with length-calibrated token reweighting. LC-DPO applies a reweighting mask only to the longer response: tokens with larger policy–reference loss gaps are preserved as preference-relevant signals, while lower-gap extra tokens are softly downweighted. $\tilde{\Delta} = \log R_\pi(y_w \mid x) - \log R_\pi(y_l \mid x)$ means weighted log-likelihood ratios, where $R_\pi(y \mid x) = \pi_\theta(y \mid x)/\pi_{\text{ref}}(y \mid x)$.

3) Training Details: We used the same hyperparameters for all methods: a learning rate of $4e$-6, a batch size of 32, and 2 epochs. Unless otherwise stated, we set $\alpha = 0.5$ in all experiments. All remaining hyperparameters and decoding settings are kept identical across all methods. All experiments were conducted on 8 × NVIDIA H20 GPUs. We repeated key experiments and observed consistent qualitative trends.

IV. Results and Discussion

In this section, we present our main experimental results, ablation study, parameter study and case study.

A. Main Results

Table I summarizes results on MT-Bench and AlpacaEval2 for three backbones (Llama3-8B-Instruct, Mistral-8B-Instruct, and Qwen2.5-1.5B-Instruct) trained on UltraFeedback. Across all models, LC-DPO delivers a consistently better length–quality trade-off than standard DPO and other length-control baselines. Compared to DPO, LC-DPO reduces the average response length in all settings (e.g., MT-Bench average tokens decrease by roughly 13%–18%, and AlpacaEval2 average tokens decrease by roughly 18%–19%), while maintaining or improving alignment scores.

To further evaluate general instruction-following and knowledge-intensive generation, we report performance on six widely used benchmarks in Table II. We evaluate Llama3-8B-Instruct and Qwen2.5-7B-Instruct under the same protocol and compare LC-DPO against several preference alignment baselines. LC-DPO achieves the best overall average score on both backbones, showing that its gains extend beyond pairwise evaluation benchmarks.

Finally, to assess whether the observed improvements are specific to UltraFeedback or extend to other preference corpora, we conduct an additional cross-dataset evaluation on HH-RLHF [33] in the next subsection.

B. Results on HH-RLHF

To assess the generalization of LC-DPO, we conduct a lightweight evaluation on a subset of the HH-RLHF dataset using Llama3-8B-Instruct. We compare DPO and LC-DPO under identical training settings. This experiment serves as a minimal cross-dataset validation rather than an exhaustive benchmark. As shown in Table III, LC-DPO consistently reduces response length while maintaining or improving alignment metrics, indicating that the proposed token-level reweighting generalizes across different preference datasets.

C. Ablation Study

1) Token selection rule Random vs Loss Only vs Gap: To examine whether the effectiveness of LC-DPO stems merely from reducing the effective number of tokens or from the specific token selection principle, we compare our loss-gap based selection with two alternatives: *Random*, which randomly select the same number of tokens on the longer response, and *Loss Only*, which select tokens solely based on token-level loss magnitude. As shown in Table V, random selection reduces the average response length but leads to a noticeable degradation on MT-Bench, indicating that uniformly discarding tokens is insufficient for preserving preference quality. Selecting tokens by loss magnitude alone yields comparable length reduction but cannot consistently improve alignment metrics.

In contrast, the proposed loss-gap based selection achieves the best overall trade-off, improving MT-Bench score while substantially increasing AlpacaEval2 LC and maintaining shorter responses. This result offers an intuitive explanation of LC-DPO: the gain does not come from uniformly suppressing longer responses, but from selectively downweighting lower-gap tokens that are less informative for preference learning.

2) Apply-side ablation chosen vs rejected vs longer-only: A potential source of length bias in DPO is the *unbalanced token mass* contributed by the longer response within each preference pair. Importantly, the longer side is not always

TABLE I

RESULTS ON MT-BENCH AND ALPACAEVAL2 UNDER DIFFERENT SETTINGS. SCORE DENOTES THE AVERAGE RATING OF MODEL RESPONSES EVALUATED BY GPT-4-TURBO-0409. LC(%) INDICATES THE LENGTH-CONTROLLED WIN RATE AGAINST THE BASELINE GPT-4-1106-PREVIEW, WHICH MITIGATES LENGTH BIAS. AVG. TOKEN DENOTES THE AVERAGE RESPONSE LENGTH. WE BOLD THE BEST RESULTS AND UNDERLINE THE SECOND BEST RESULTS.

Methods	Llama3-8B-Instruct				Mistral-8B-Instruct				Qwen2.5-1.5B-Instruct			
	MT-Bench		AlpacaEval2		MT-Bench		AlpacaEval2		MT-Bench		AlpacaEval2	
	Score	Avg. Token	LC (%)	Avg. Token	Score	Avg. Token	LC (%)	Avg. Token	Score	Avg. Token	LC (%)	Avg. Token
SFT	6.91	309	34.92	324	7.55	240	29.53	276	5.88	249	19.34	271
DPO	7.25	372	39.28	393	7.72	297	34.15	332	6.50	296	21.57	340
R-DPO	6.88	**303**	37.96	**318**	7.56	**242**	32.57	271	6.06	268	20.39	**276**
SimPO	7.12	335	39.65	353	7.73	267	34.73	284	6.38	286	22.83	304
SamPO	7.06	319	40.77	332	7.75	251	**35.11**	**264**	6.42	263	21.39	286
LD-DPO	7.23	341	40.07	357	7.70	263	34.61	279	6.24	271	23.05	294
LC-DPO	**7.31**	308	**42.07**	321	**7.79**	244	35.07	272	**6.58**	**259**	**23.91**	279

TABLE II

QUANTITATIVE RESULTS OF FINE-TUNING LLAMA3-8B-INSTRUCT AND QWEN2.5-7B-INSTRUCT WITH DPO, SEVERAL VARIANTS AND OUR LC-DPO. WE EVALUATE ALL MODELS UNDER SAME FRAMEWORK.

Methods	LLama3-8B-Instruct							Qwen2.5-7B-instruct						
	GSM8K	IFEval	PiQA	MMLU	COQA	HellaSwag	Avg.	GSM8K	IFEval	PiQA	MMLU	COQA	HellaSwag	Avg.
SFT	80.13	55.63	78.45	65.86	61.33	57.66	66.51	83.16	72.30	78.48	73.73	57.23	77.05	73.66
DPO	80.57	59.03	76.27	66.51	61.83	57.92	67.02	82.54	73.34	76.27	72.93	56.17	76.86	73.02
R-DPO	79.89	56.92	76.57	65.48	61.54	57.79	66.37	81.41	71.86	77.82	71.35	57.21	76.83	72.75
SimPO	79.64	61.41	77.86	65.79	61.41	59.41	67.59	83.82	71.38	79.81	73.77	56.45	76.98	73.70
SamPO	82.48	59.26	78.69	66.74	**63.97**	58.96	68.35	**84.96**	74.76	**80.49**	72.86	**59.64**	76.71	74.90
LD-DPO	81.32	**62.08**	78.65	66.71	61.04	57.04	67.81	84.03	72.05	79.92	73.81	58.43	78.54	74.46
LC-DPO	**82.86**	60.39	**78.78**	**66.89**	63.15	**59.47**	**68.59**	84.94	**74.81**	80.47	**73.85**	57.31	**79.02**	**75.07**

TABLE III

RESULTS ON HH-RLHF USING LLAMA3-8B-INSTRUCT.

Method	MT-Bench		AlpacaEval2	
	Score	Avg. Token	LC (%)	Avg. Token
SFT	6.91	309	34.92	324
DPO	7.17	359	37.83	406
LC-DPO	**7.28**	314	**40.76**	**319**

TABLE IV

ABLATION STUDY ON LLAMA3-8B-INSTRUCT: CHOSEN DENOTES LENGTH DECOUPLING APPLIED ONLY TO y_w, REJECTED DENOTES ONLY TO y_l.

Method	MT-Bench		AlpacaEval2	
	Score	Avg. Token	LC (%)	Avg. Token
DPO	7.25	372	39.28	393
Chosen	7.16	312	40.36	348
Rejected	7.13	315	41.58	335
LC-DPO	**7.31**	**308**	**42.07**	**321**

the chosen response: depending on the dataset, either y_w or y_l may contain extra tokens beyond the common length T_p. To diagnose *which side primarily contributes* to the observed length bias and test whether length calibration must be applied symmetrically, we consider controlled variants that apply our reweighting only to y_w, only to y_l, or to both responses.

As shown in Table IV, applying our reweighting strategy to only the chosen or only the rejected response provides some degree of length control. However, neither partial variant matches full LC-DPO, indicating that jointly reweighting both sides better mitigates length-related bias in pairwise preference learning, and achieves the best overall training performance.

D. Parameter Study

We study the effect of the hyperparameter α, which controls the strength of token downweighting on the longer response and thus the degree of length–preference decoupling. When $\alpha = 1$, LC-DPO reduces to standard DPO and decreasing α progressively downweights extra tokens beyond the common length, enforcing stronger length calibration.

Fig. 5 shows that the average response length decreases as α becomes smaller. Notably, the rate of reduction slows once α drops below approximately 0.5, suggesting that most length-induced bias has already been mitigated by this point.

In terms of alignment quality, performance on MT-Bench and AlpacaEval2 shows a non-monotonic trend: metrics first improve as α decreases from 1.0, peak around 0.5, and then gradually decline when α becomes too small. This reflects a trade-off between correcting length bias and preserving informative tokens for preference learning. Overall, a moderate α achieves the best balance between length control and preference optimization.

V. CONCLUSION

In this paper, we introduced LC-DPO, a simple length-controlled variant of DPO that mitigates verbosity bias while preserving alignment quality. By assigning dynamic token weights based on policy–reference loss gaps, LC-DPO em-

TABLE V
ABLATION ON TOKEN SELECTION RULES (SAME TOKEN BUDGET T_p ON THE LONGER RESPONSE): *Random*, *Loss Only*, AND OURS *Gap*.

Select Rule	MT-Bench		AlpacaEval2	
	Score	Avg. Token	LC (%)	Avg. Token
DPO	7.25	372	39.28	393
LC-DPO(Random)	7.11	307	39.64	323
LC-DPO(Loss Only)	7.09	310	38.91	326
LC-DPO(Gap, ours)	**7.31**	308	**42.07**	**321**

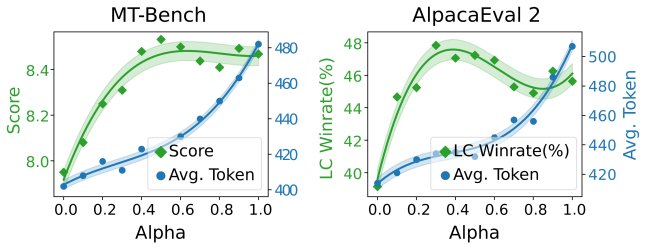

Fig. 5. Hyperparameter analysis on α with Qwen2.5-7B-Instruct.

phasizes preference-relevant tokens and downweights less informative verbose ones. Across multiple backbones and benchmarks, LC-DPO consistently improves the length–quality trade-off over standard DPO and representative baselines, reducing average response length by 13%–20%. Beyond its empirical gains, LC-DPO is particularly useful in deployment scenarios where unnecessary verbosity affects user experience and serving cost. More broadly, our results suggest that token level calibration within preference pairs is a promising direction for decoupling response quality from response length in preference alignment.

ACKNOWLEDGMENT

This research work was supported by the Tianchi Talents Introduction Plan and the Shaanxi Provincial Key R&D Program (Project No.: 2025CY-JJQ-140).

REFERENCES

[1] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, et al., "A survey on evaluation of large language models," ACM transactions on intelligent systems and technology, vol. 15, no. 3, pp. 1–45,2024.
[2] Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, et al., "A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment," arXiv preprint arXiv:2504.15585, 2025.
[3] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, et al., "Hybridflow: A flexible and efficient rlhf framework," in EuroSys, 2025, pp. 1279–1297.
[4] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Zhiyuan Liu, et al., "Ultrafeedback: Boosting language models with high-quality feedback, 2024," in URL https://openreview. net/forum, 2023.
[5] Zixiang Chen, Yihe Deng, Huizhuo Yuan, et al., "Self-play fine-tuning converts weak language models to strong language models, 2024," URL https://arxiv. org/abs/2401.01335, 2024.
[6] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu, "A comprehensive survey of reward models: Taxonomy, applications, challenges, and future," arXiv preprint arXiv:2504.12328, 2025.
[7] Leo Gao, John Schulman, and Jacob Hilton, "Scaling laws for reward model overoptimization," in International Conference on Machine Learning. PMLR, 2023, pp. 10835–10866.
[8] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov, "Proximal policy optimization algorithms," arXiv preprint rXiv:1707.06347, 2017.
[9] Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, et al., "Secrets of rlhf in large language models part ii: Reward modeling," arXiv preprint arXiv:2401.06080, 2024.
[10] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, et al., "Direct preference optimization: Your language model is secretly a reward model," Advances in neural information processing systems, vol. 36, pp. 53728–53741, 2023.
[11] Zhenyu Hou, Yilin Niu, Zhengxiao Du, Xiaohan Zhang, Xiao Liu, et al., "Chatglm-rlhf: Practices of aligning large language models with human feedback," arXiv preprint arXiv:2404.00934, 2024.
[12] Wei Zhang, Zhenhong Zhou, Kun Wang, Junfeng Fang, Yuanhe Zhang, et al., "Lifebench: Evaluating length instruction following in large language models," arXiv preprint arXiv:2505.16234, 2025.
[13] 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, 2024.
[14] Wei Liu, Yang Bai, Chengcheng Han, Rongxiang Weng, Jun Xu, et al., "Length desensitization in direct preference optimization," 2024.
[15] Gengxu Li, Tingyu Xia, Yi Chang, and Yuan Wu, "Length-controlled margin-based preference optimization without reference model," arXiv preprint arXiv:2502.14643, 2025.
[16] Yu Meng, Mengzhou Xia, and Danqi Chen, "Simpo: Simple preference optimization with a reference-free reward," Advances in Neural Information Processing Systems, vol. 37, pp.124198–124235, 2024.
[17] Yuyan Bu, Liangyu Huo, Yi Jing, and Qing Yang, "Beyond excess and deficiency: Adaptive length bias mitigation in reward models for rlhf," in NAACL, 2025, pp. 3091–3098.
[18] Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Di Yin, et al., "Eliminating biased length reliance of direct preference optimization via down-sampled kl divergence," arXiv preprint arXiv:2406.10957, 2024.
[19] Aiwei Liu, Haoping Bai, Zhiyun Lu, Yanchao Sun, et al., "Tis-dpo: Token-level importance sampling for direct preference optimization with estimated weights," arXiv preprint arXiv:2410.04350, 2024.
[20] Ning Yang, Hai Lin, Yibo Liu, Baoliang Tian, Guoqing Liu, et al., "Token-importance guided direct preference optimization," arXiv preprint arXiv:2505.19653, 2025.
[21] Meng Li, Guangda Huzhang, Haibo Zhang, Xiting Wang, and Anxiang Zeng, "Optimal transport-based token weighting scheme for enhanced preference optimization," arXiv preprint arXiv:2505.18720, 2025.
[22] AI@Meta, "Llama 3 model card," 2024.
[23] Qwen Team et al., "Qwen2 technical report," arXiv preprint arXiv:2407.10671, 2024.
[24] Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, et al., "Mt-bench-101: A finegrained benchmark for evaluating large language models in multi-turn dialogues," arXiv preprint arXiv:2402.14762, 2024.
[25] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, et al., "Alpacaeval: An automatic evaluator of instruction-following models," https://github.com/tatsu-lab/alpaca_eval,5 2023.
[26] Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, et al., "Rho-1: Not all tokens are what you need," arXiv preprint arXiv:2404.07965, 2024.
[27] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, et al., "Training verifiers to solve math word problems," arXiv preprint arXiv:2110.14168, 2021.
[28] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, et al., "Instruction-following evaluation for large language models," arXiv preprint arXiv:2311.07911, 2023.
[29] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, "Piqa: Reasoning about physical commonsense in natural language," in Proceedings of the AAAI conference on artificial intelligence, 2020, vol. 34, pp. 7432–7439.
[30] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, et al., "Measuring massive multitask language understanding," arXiv preprint arXiv:2009.03300, 2020.
[31] Siva Reddy, Danqi Chen, and Christopher D Manning, "Coqa: A conversational question answering challenge," Transactions of the Association for Computational Linguistics, vol. 7, pp. 249–266, 2019.
[32] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi, "Hellaswag: Can a machine really finish your sentence?," arXiv preprint arXiv:1905.07830, 2019.
[33] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, et al. "Training a helpful and harmless assistant with reinforcement learning from human feedback," arXiv preprint arXiv:2204.05862, 2022.