LENGTH IS NOT JUST BIAS: MARGINAL TOKEN UTILITY IN LLM POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

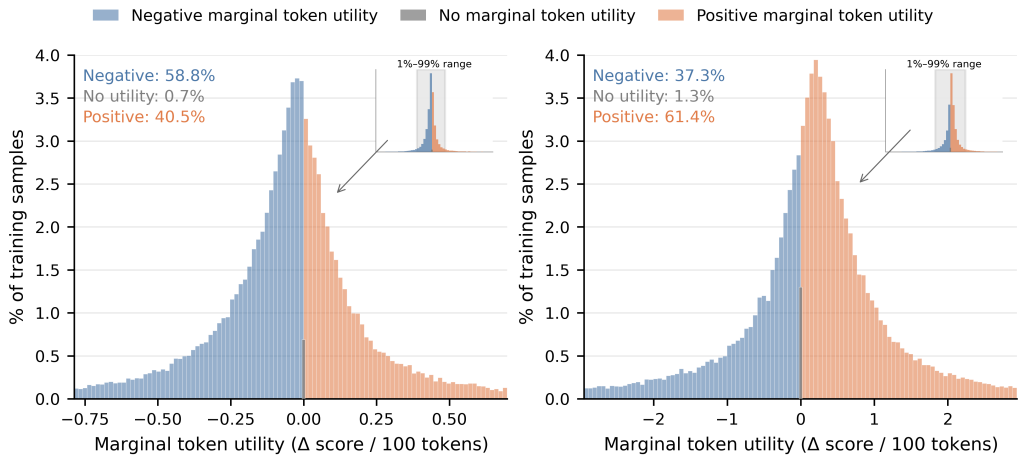
Response length is often treated as a source of bias in LLM post-training, motivating methods that globally suppress or normalize its influence on learning. We study this issue in preference-based post-training, where our analysis of training data shows that the relationship between response length and quality may not be universal: additional length is associated with better outcomes in some contexts, contributes little in others, and can be detrimental elsewhere. To capture this context-dependent relationship, we introduce marginal token utility, which measures the gain associated with additional response tokens within each context. This perspective suggests that length should not be uniformly penalized or removed from optimization. Instead, its influence should reflect the marginal gains associated with additional tokens. Building on this insight, we propose a marginal-utility-guided objective that modulates length-related learning according to marginal token utility. Experiments across language models and alignment benchmarks show that our approach consistently outperforms existing length-debiasing methods while avoiding unnecessary verbosity. Our results suggest that response length should be viewed not merely as an optimization bias, but as a context-dependent source of marginal utility in LLM post-training.

1 INTRODUCTION

Post-training has become a central approach for aligning large language models (LLMs) with human preferences and improving response quality (Ouyang et al., 2022; Rafailov et al., 2023). Yet preference and reward signals, as well as automatic evaluations, can become entangled with superficial response characteristics, among which response length is particularly important (Singhal et al., 2024; Dubois et al., 2024; Feuer et al., 2025; Huang et al., 2025). Models may exploit correlations between response length and these alignment signals, producing increasingly verbose responses as a shortcut to higher reward. These extra tokens, however, may provide little corresponding improvement in quality, a phenomenon commonly characterized as *length bias* or *reward hacking* (Singhal et al., 2024; Park et al., 2024; Lu et al., 2024). Existing approaches largely address this issue in two ways: by directly suppressing excessive length through regularization or length penalties, or by reducing the dependence of learning signals on response length through token-level balancing and reward-model debiasing (Park et al., 2024; Chen et al., 2025; Lu et al., 2024; Huang et al., 2025). However, these approaches largely apply a universal treatment of response length across contexts, either discouraging additional tokens or reducing their influence regardless of the value they provide.

Recent work offers initial evidence that this universal length treatment may be too restrictive: longer responses may contain additional useful information (Hu et al., 2025), and legitimate length preferences can vary across queries (Bu et al., 2025). Our analysis of training data reveals a more fine-grained, context-dependent pattern: additional response tokens are associated with positive quality gains in some contexts, little measurable gain in others, and negative gains elsewhere. As shown in Fig. 1, marginal token utility spans a broad range of both positive and negative values within each training setting, indicating substantial variation across individual training samples. Both directions account for a considerable fraction of the data: 58.8% negative versus 40.5% positive in the same-model setting, and 37.3% negative versus 61.4% positive in the cross-model setting. The aggregate tendency also differs across settings, which may reflect differences in their response sources. A dataset-level length–quality statistic would average over these opposing context-level patterns, yielding a weaker overall trend. Consequently, a universal treatment of length cannot distinguish

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.



(a) Same-model training data

(b) Cross-model training data

Figure 1: Marginal token utility varies substantially across samples, with substantial positive and negative values in both settings, motivating adaptive rather than universal length control. We show the 1st–99th percentile range for clarity due to long-tailed distribution extremes. Details of the same-model and cross-model training data settings are provided in Section 5.1.

additional tokens that contribute meaningful improvements from those that provide little value or introduce unnecessary verbosity.

To characterize this variation, we introduce *marginal token utility*, defined as the gain in training score associated with additional response length within a given context. Positive marginal token utility indicates that additional tokens are associated with improved response quality, near-zero utility indicates little return from further length, and negative utility indicates that additional length is associated with worse outcomes. This perspective shifts the question from whether response length should generally be suppressed or removed to *when additional token usage is worthwhile*. In this sense, *length is not just a universal bias*: additional tokens may reflect undesirable verbosity in some contexts while providing worthwhile gains in others.

Building on this perspective, we propose *UTIL*, which uses marginal token utility to adaptively control the role of response length for each context during post-training. We estimate marginal token utility from the score structure of each training context and use it to adapt how response length contributes during post-training. The resulting objective is bidirectional: it favors shorter responses when additional tokens have negative utility, remains close to length-neutral when their utility is negligible, and preserves or favors additional generation when their utility is positive. We instantiate UTIL based on direct preference optimization (Rafailov et al., 2023), a representative and computationally efficient approach to preference-based post-training, replacing globally applied length control with context-dependent control guided by marginal token utility. This design allows post-training to reduce unnecessary token use while retaining additional generation where it contributes to response quality.

We evaluate UTIL across multiple language models, training-data settings, and standard alignment benchmarks. Our results show consistent improvements over representative length-bias mitigation methods while avoiding unnecessary verbosity. Additional analysis shows that uniform treatments of response length produce different quality–efficiency trade-offs across settings, whereas marginal token utility enables this trade-off to adapt according to the value of additional generation. Together, these results suggest that effective length control in post-training should account not only for the cost of additional tokens, but also for the quality gains they provide.

Our main contributions are:

- We show that the quality gain associated with additional tokens is heterogeneous across contexts, revealing diverse length–quality associations hidden behind aggregate statistics.

- We introduce *marginal token utility* to characterize the context-dependent value of additional response tokens, reframing length control around the trade-off between response quality and token usage rather than treating length solely as a bias or reward hacking.
- We propose UTIL, a token-utility-guided method that adaptively modulates the role of response length during post-training according to its context-specific utility, improving over existing length-bias mitigation methods across models and benchmarks by up to 7.7% in Win Rate and 4.6% in LC Win Rate.

2 Related Work

LLM length bias and reward-model debiasing. Response length has long been recognized as a potentially confounding factor in LLM alignment and evaluation (Zheng et al., 2023). Prior work finds that increases in reward are often strongly associated with longer responses (Singhal et al., 2024), while LLM-based evaluators can favor stylistic characteristics over substantive response quality (Feuer et al., 2025). More broadly, optimizing imperfect reward signals can lead to reward overoptimization (Gao et al., 2023). These observations motivate a broad line of work that treats verbosity and related stylistic properties as shortcut features in reward or preference modeling (Zhang et al., 2025), with mitigation approaches including post-hoc reward calibration (Huang et al., 2025; Zhao et al., 2026), robust reward-model training (Liu et al., 2025), shortcut-aware reward learning (Ye et al., 2025), and counterfactual preference construction (Bharadwaj et al., 2026). Overall, this line of work primarily addresses length dependence on the *signal side* of alignment, by correcting reward models, training data, or evaluators so that response length is less likely to act as a spurious proxy for quality.

Length-aware post-training. Several recent studies have examined how response length should be handled during model post-training. R-DPO introduces a length-dependent regularization term into direct preference optimization, with the trade-off between response quality and length controlled by a global coefficient (Park et al., 2024). SamPO instead reduces biased reliance on response length through token-level balancing based on down-sampled KL divergence (Lu et al., 2024). Length control has also been incorporated into broader post-training pipelines: DICE uses length-regularized implicit rewards when constructing iterative preference data (Chen et al., 2025), while Rc-DPO combines response-conditioned reward modeling, data construction, and policy optimization to distinguish spurious length bias from legitimate response-length requirements (Cai et al., 2026). These methods largely control length through regularization, normalization, or data construction, without explicitly considering the value of additional generation. Such treatments can also lead to undesirable effects such as premature termination (Gupta et al., 2025), motivating a more context-aware treatment of response length.

Context-dependent effects of length. A limited line of work has explored whether the value of response length itself may vary across contexts. Longer responses can sometimes be preferred because they contain additional useful information rather than merely reflecting superficial length bias (Hu et al., 2025), and reward-model studies have shown that legitimate length preferences can vary across queries (Bu et al., 2025). However, these observations have mainly been studied in evaluation or reward modeling. How to characterize the *marginal value of additional tokens* within each context and use it directly to guide policy post-training remains underexplored. Our work addresses this gap through marginal token utility, which quantifies the gain associated with additional response tokens and uses it to adapt how response length contributes during post-training.

3 Preliminaries

Post-training adapts language models beyond pretraining and supervised instruction tuning using feedback about desirable model behavior. Such feedback may come from human annotations, reward models, or stronger language models, and can provide either scalar quality scores or relative judgments among candidate responses. Given a context x and response y, we use $r(y \mid x)$ to denote its training quality signal. Our method builds on pairwise preference optimization, where candidate responses are compared according to their relative quality. For a context x, let y_w and y_l denote the preferred and rejected responses, with lengths $L_w = L(y_w)$ and $L_l = L(y_l)$, respectively. We

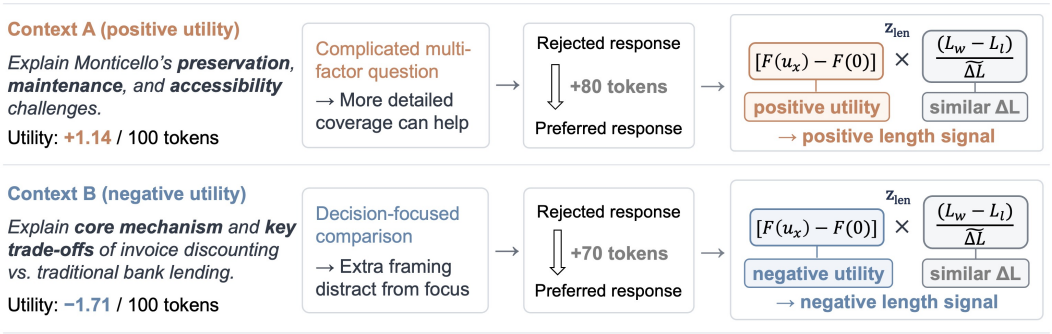


Figure 2: Illustration of UTIL's context-dependent length signal. Preference pairs can exhibit similar response-length differences while having different marginal token utility across contexts. UTIL combines the context-specific utility with the observed length difference, allowing additional tokens to contribute differently according to the value of additional generation.

instantiate our method with Direct Preference Optimization (DPO) (Rafailov et al., 2023), a representative and computationally efficient approach to preference-based post-training. DPO increases the relative likelihood of the preferred response with respect to a reference policy through the pairwise preference logit

$$z_{\mathrm{DPO}} = \beta \left[\log \frac{\pi_\theta(y_w \mid x)}{\pi_{\mathrm{ref}}(y_w \mid x)} - \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\mathrm{ref}}(y_l \mid x)} \right], \qquad \mathcal{L}_{\mathrm{DPO}} = -\log \sigma(z_{\mathrm{DPO}}). \tag{1}$$

Response length can also influence preference optimization when candidate responses differ in the amount of generated content. Existing approaches largely address this effect through globally applied length penalties, normalization, or controls. Yet the value of additional generation can vary across contexts: extra tokens may improve response quality in some cases, provide little additional value in others, or even be detrimental. This variation motivates our context-dependent treatment of response length through marginal token utility.

4 Method

We propose *UTIL*, which treats response length according to the context-specific marginal value of additional generation rather than through a universal preference for shorter or longer responses. The key idea is *marginal token utility*, which characterizes how response quality changes with additional tokens for each context. As illustrated in Fig. 2, two preference pairs can exhibit similar response-length differences while the value of additional generation differs substantially across contexts. UTIL combines this context-specific utility with the observed preferred–rejected length difference, allowing the length-related learning signal to vary according to whether additional generation is useful for the current context.

4.1 Marginal Token Utility

The value of additional generation can vary substantially across contexts: additional tokens may provide useful reasoning, supporting evidence, or clarification in some cases, while offering little additional value or even reducing quality in others. A raw difference in response length therefore does not by itself indicate whether the additional generation is worthwhile. We propose marginal token utility to characterize this context-dependent value of additional generation.

Definition 1 (**Marginal Token Utility**). Let $Y \sim \pi(\cdot \mid x)$ denote a response generated by policy π for context x, and let $L(Y)$ denote its length measured by the number of generated tokens. At a given token length ℓ, the conditional expectation $\mathbb{E}_{Y \sim \pi(\cdot|x)} [r(Y \mid x) \mid L(Y) = \ell]$ describes the expected response quality associated with that amount of generation. We define marginal token utility as the change in expected response quality associated with an incremental increase in token count:

$$u_x(\ell) = \frac{\partial \, \mathbb{E}_{Y \sim \pi(\cdot|x)} [r(Y \mid x) \mid L(Y) = \ell]}{\partial \ell}. \tag{2}$$

This definition directly captures the marginal return of additional generation. Positive utility indicates that generating additional tokens is associated with higher expected quality, values near zero indicate little measurable return from further generation, and negative utility indicates that additional tokens are associated with lower expected quality. Importantly, this dependence is conditioned on the context x: the same increase in token usage can therefore carry very different, or even opposite, value across prompts.

LLM post-training often collects multiple generated candidate responses for each context and evaluates them using quality scores or annotations. For a context x, let $\{y_j\}_{j=1}^{m_x}$ denote these candidates, with quality scores $r(y_j \mid x)$. Their variation in token length and quality provides empirical observations of the within-context quality–length dependence. We approximate Eq. 2 from these finite observations by estimating the average rate of quality change across the available candidates:

$$u_x = \frac{\sum_j \left(L(y_j) - \bar{L}_x\right)\left(r(y_j \mid x) - \bar{r}_x\right)}{\sum_j \left(L(y_j) - \bar{L}_x\right)^2}, \tag{3}$$

where $\bar{L}_x$ and $\bar{r}_x$ denote the mean token length and quality score, respectively, among candidate responses for x. This estimate captures the average marginal change in quality associated with additional tokens over the observed responses. Higher scores for longer responses yield positive utility, while lower scores as length increases yield negative utility.

The resulting u_x preserves the context-specific dependence between response quality and generation length for each x rather than collapsing it into an aggregate dataset-level tendency. A single corpus-level statistic can obscure contexts in which additional tokens have fundamentally different returns. As shown in Fig. 1, the empirical distribution of marginal token utility spans substantial positive and negative regions even within the same training setting. Additional generation is therefore neither uniformly beneficial nor uniformly undesirable; its value depends on the context in which those tokens are used. UTIL carries this context-specific signal into post-training so that the role of response length reflects the quality return associated with additional generation.

4.2 Token-Utility-Guided Post-Training

Marginal token utility provides a context-specific signal for how response length contributes to preference optimization. Building on the pairwise preference signal z_{DPO} in Eq. 1, UTIL separates response quality from the effect of response-length differences, allowing both to be represented explicitly in the objective. The resulting preference signal is written as

$$z_{\mathrm{UTIL}} = z_{\mathrm{qual}} + z_{\mathrm{len}}, \tag{4}$$

where z_{qual} captures the length-neutral quality signal, while z_{len} captures how the response-length difference contributes according to its context-specific marginal token utility.

Length-neutral quality signal. UTIL first captures the quality preference signal using comparable amounts of generation from the preferred and rejected responses. This provides a common basis for the quality comparison, while their difference in generation length is represented separately through the token-utility-guided length signal. Specifically, we uniformly sample token positions from each response (Lu et al., 2024), denoted by $\mathcal{I}_w$ and $\mathcal{I}_l$, with $|\mathcal{I}_w| = |\mathcal{I}_l| = \min(L_w, L_l)$. The quality signal is then

$$z_{\mathrm{qual}} = \beta \left[\sum_{t \in \mathcal{I}_w} \log \frac{\pi_\theta(y_{w,t} \mid y_{w,<t}, x)}{\pi_{\mathrm{ref}}(y_{w,t} \mid y_{w,<t}, x)} - \sum_{t \in \mathcal{I}_l} \log \frac{\pi_\theta(y_{l,t} \mid y_{l,<t}, x)}{\pi_{\mathrm{ref}}(y_{l,t} \mid y_{l,<t}, x)} \right]. \tag{5}$$

Letting an equal number of tokens contribute from both responses keeps z_{qual} focused on the quality preference signal. The remaining difference in response length is then captured explicitly through z_{len}, as determined by marginal token utility.

Token-utility-guided length signal. In pairwise preference optimization, the preferred and rejected responses naturally exhibit a difference in generation length. Specifically, $L_w - L_l$ indicates whether, and by how much, the preferred response is longer or shorter than the rejected response. We normalize the pairwise length difference by its typical magnitude in the training data so that

differences across pairs are represented on a comparable scale: $\widetilde{\Delta L} = \text{median}_{(w,l)} |L_w - L_l|$. The normalized term $(L_w - L_l)/\widetilde{\Delta L}$ captures the relative amount of additional generation in the current preference pair.

However, this length difference alone does not indicate whether the additional generation is useful for the context. Marginal token utility u_x provides this context-specific interpretation by characterizing the quality return of additional generation for context x. For example, the same large positive length difference can strengthen the preference signal when additional generation has positive utility, but weaken it when the utility is negative.

To compare the strength of this utility across training samples, we represent u_x by its relative position in the overall utility distribution. Let F denote the empirical cumulative distribution function (CDF) of u_x over the training set, where $F(u_x)$ is the fraction of contexts whose utility is no greater than u_x. Measuring this position relative to the neutral utility $u_x = 0$ gives $F(u_x) - F(0)$, which preserves the direction of marginal token utility while reflecting its relative distance from neutral utility in the training distribution.

This leads to a natural construction: the observed pairwise length difference is weighted by the context-specific marginal utility of additional generation. In this way, the magnitude of the length difference determines how much additional generation is involved, while marginal token utility determines how that difference is interpreted for the current context. We therefore define the token-utility-guided length signal as

$$z_{\text{len}} = [F(u_x) - F(0)] \frac{L_w - L_l}{\widetilde{\Delta L}}. \tag{6}$$

This formulation turns the observed length difference into a context-aware contribution to preference optimization. Combining the quality signal and the token-utility-guided length signal gives the final optimization objective:

$$\mathcal{L}_{\text{UTIL}} = -\log \sigma \left(z_{\text{qual}} + z_{\text{len}} \right). \tag{7}$$

The objective preserves the quality preference signal while allowing the contribution of response-length differences to vary according to their context-specific marginal token utility. The signed z_{len} can either reinforce or counteract the quality signal according to the context-specific token utility and observed length difference.

5 Experiments

5.1 Experimental Setup

Models and Training Data. We evaluate UTIL under two complementary and representative preference-training settings based on UltraFeedback (Cui et al., 2024), differing in how closely the training responses are matched to the target model. In the same-model setting, we use Llama-3 UltraFeedback-ArmoRM[1], where candidate responses are generated by Llama-3-8B-Instruct and scored by ArmoRM, and train Llama-3-8B-Instruct as the target model. In the cross-model setting, we use UltraFeedback[2], which contains responses generated by multiple language models, and train Qwen2.5-7B-Instruct. The former provides model-matched preference data, while the latter represents a more heterogeneous and reusable training regime. Together, the two settings allow us to evaluate UTIL across both different model families and different degrees of correspondence between the training responses and the target model.

Baselines. This paper focuses on representative methods that address length dependence from different aspects, rather than comparing against a broad collection of generic preference optimization variants. DPO (Rafailov et al., 2023) serves as the standard pairwise preference-training baseline. R-DPO (Park et al., 2024) introduces a global response-length regularization term, representing explicit and uniform pressure on response length. SamPO (Lu et al., 2024) instead reduces biased length reliance by balancing token-level contributions through down-sampled KL divergence. ConfPO (Yoon et al., 2025) selectively emphasizes preference-critical tokens according to policy

[1] https://huggingface.co/datasets/princeton-nlp/llama3-ultrafeedback-armorm

[2] https://huggingface.co/datasets/openbmb/UltraFeedback

Table 1: Results on Arena-Hard and AlpacaEval 2. We report win rate (Win Rate), length-controlled win rate (LC Win Rate), and average response length in tokens (Avg. Tokens). The base model performance is shown for reference. Among post-training methods, the best result is shown in **bold** and the second-best is underlined.

Method	Arena-Hard			AlpacaEval 2		
	Win Rate ↑	LC Win Rate ↑	Avg. Tokens	Win Rate ↑	LC Win Rate ↑	Avg. Tokens
Llama-3-8B-Instruct (*same-model training data*)						
Base	38.40	39.90	572	31.42	32.12	423
DPO	48.30	50.80	567	<u>36.59</u>	36.97	430
R-DPO	33.50	46.20	444	31.64	37.03	370
SamPO	<u>49.00</u>	<u>53.30</u>	547	36.44	<u>37.79</u>	417
ConfPO	37.10	40.40	570	32.25	33.18	425
UTIL	**52.60**	**54.90**	583	**39.41**	**39.54**	435
Qwen2.5-7B-Instruct (*cross-model training data*)						
Base	74.50	70.60	703	37.48	39.81	411
DPO	<u>73.60</u>	69.60	688	36.04	39.43	397
R-DPO	71.00	69.90	617	34.28	<u>42.45</u>	352
SamPO	73.30	<u>71.50</u>	676	<u>39.04</u>	41.91	395
ConfPO	72.60	71.10	662	36.26	41.12	376
UTIL	**75.80**	**73.60**	675	**39.37**	**42.73**	397

confidence, providing a complementary token-selective treatment of the preference signal without imposing a global length penalty. Together, these baselines cover global length regularization, token-level length balancing, and selective token-level learning, providing broad and complementary coverage for comparison with UTIL's context-dependent treatment of response length. All compared methods within each setting use the same backbone, training data, and training configuration. Full setting details are provided in Appendix A.

Benchmarks and Metrics. We evaluate on Arena-Hard (Li et al., 2025) and AlpacaEval 2 (Dubois et al., 2024), two widely used benchmarks for instruction-following and alignment evaluation. For both benchmarks, we report win rate (Win Rate), length-controlled win rate (LC Win Rate), and average response length in tokens. Win Rate measures pairwise preference performance, while LC Win Rate accounts for evaluator sensitivity to response length. Average response length further characterizes the amount of generation produced by each method. Together, these metrics allow us to examine alignment quality alongside the quality–length trade-off.

5.2 Main Results

Overall Performance. Table 1 shows that UTIL consistently achieves the highest Win Rate and LC Win Rate across both models and both benchmarks, demonstrating gains under both conventional and length-controlled preference evaluation. The improvement is particularly strong for Llama-3-8B-Instruct with same-model training data: on AlpacaEval 2, UTIL improves over the strongest competing baseline by 7.7% in Win Rate and 4.6% in LC Win Rate. Similar improvements are observed across the remaining settings, with UTIL ranking first on both metrics throughout. These consistent gains support the central design of UTIL: adaptive learning guided by context-specific marginal token utility preserves useful additional generation while reducing the influence of low-utility generation, compared with a uniform treatment of response length.

Quality–Token Trade-off. Figure 3 further shows that the performance gains do not arise from systematically increasing or decreasing the overall amount of generation. Apart from R-DPO, which explicitly applies global length reduction, the compared preference-training methods operate within broadly similar average response-length ranges, yet UTIL consistently achieves the strongest Win Rate and LC Win Rate. This suggests that the key difference is not the aggregate token budget itself, but how additional generation is treated across contexts. UTIL uses a similar overall amount

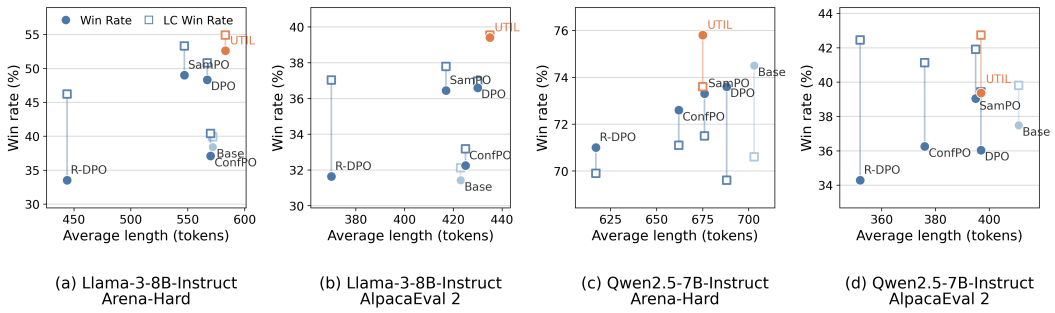


Figure 3: Quality–token trade-off across training settings and evaluation benchmarks. Win Rate and LC Win Rate are shown against average response length for each method.

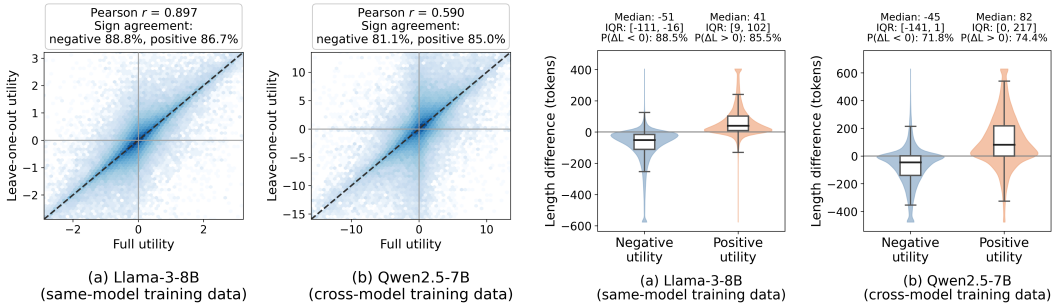


Figure 4: Analysis of marginal token utility. *Left:* Leave-one-candidate-out evaluates the robustness of utility estimates to candidate response sets. *Right:* Pairwise length differences show an overall relationship with token utility but also disagreement, especially in the heterogeneous cross-model setting, highlighting that marginal token utility captures information beyond a single preference pair.

of generation more effectively by preserving additional generation when marginal token utility is positive while applying stronger pressure against it when the utility is negative.

The comparison with R-DPO makes the limitation of universal length reduction particularly clear. R-DPO substantially shortens responses and can remain relatively competitive in LC Win Rate on AlpacaEval 2, but this often comes at the cost of actual pairwise preference wins, as reflected by its lower Win Rate. This pattern shows that aggressive length reduction can improve length-controlled evaluation while sacrificing overall preference quality. In contrast, UTIL adapts length pressure according to context-specific marginal utility, achieving stronger Win Rate and LC Win Rate without imposing a universal preference for either shorter or longer responses.

5.3 Analysis of Marginal Token Utility

Robustness of marginal token utility estimation. We examine whether marginal token utility, estimated from a finite candidate set, depends strongly on any individual response. We perform leave-one-candidate-out recomputation by removing one response from each context and comparing the resulting utility with the full-set estimate. For clarity, Figure 4 (left) reports utility as score change per 100 tokens; this rescaling is only for visualization clarity. The full and leave-one-out estimates remain strongly correlated for the same-model data ($r = 0.897$) and moderately correlated for the more heterogeneous cross-model data ($r = 0.590$). The lower correlation in the latter reflects greater variation in utility magnitude, while the overall quality–length relationship remains preserved. Utility direction is more stable: negative and positive utilities preserve their signs in 88.8% and 86.7% of recomputations for the same-model data, and 81.1% and 85.0% for the cross-model data. This matters for UTIL because the sign of marginal utility determines whether additional generation is interpreted as beneficial or detrimental, while its relative magnitude controls the strength of this interpretation. Thus, even when the exact slope is noisier, utility direction remains stable for

Table 2: Ablation study of UTIL on Llama-3-8B-Instruct.

Variant	Arena-Hard			AlpacaEval 2		
	Win Rate ↑	LC WR ↑	Avg. Tokens	Win Rate ↑	LC WR ↑	Avg. Tokens
w/o Length-Neutral Quality	48.00	51.10	576	38.00	38.08	439
w/o Utility-Guided Length	46.20	49.40	574	36.82	36.99	435
Dataset-Level Utility	49.00	53.50	559	38.02	38.39	422
UTIL	**52.60**	**54.90**	583	**39.41**	**39.54**	435

most contexts, suggesting that UTIL's adaptive signal reflects the broader candidate set rather than an individual response.

Relationship to pairwise length preference. We examine how marginal token utility relates to the observed pairwise length preference in the training data. For each context, we compute the preferred–rejected response-length difference, $\Delta L = L_w - L_l$. Figure 4 (right) shows an overall directional association: negative-utility contexts tend to favor shorter responses, whereas positive-utility contexts tend to favor longer responses. At the same time, this relationship is not deterministic. In the same-model data, 11.5% of negative utility and 14.5% of positive utility contexts deviate from the dominant direction; in the more heterogeneous cross-model setting, these proportions increase to 28.2% and 25.6%, respectively. This variation is particularly relevant because ΔL reflects only a single preferred–rejected pair, whereas marginal token utility is estimated from the broader set of candidate responses for each context. The greater disagreement in the cross-model setting may reflect its more diverse response sources, where a single pair is less representative of the overall quality–length relationship. Marginal token utility therefore provides a context-level signal that can capture information beyond the observed length difference of an individual preference pair.

5.4 Ablation Study

We conduct an ablation study on Llama-3-8B-Instruct to isolate the contribution of the main design choices in UTIL, with results reported in Table 2. All variants use the same training data and settings as the full method. (i) *w/o Length-Neutral Quality* removes the length-neutral quality construction and reduces both Win Rate and LC Win Rate, showing that separating the underlying quality preference from response-length differences provides an important complementary signal. (ii) *w/o Utility-Guided Length* removes the utility-guided length signal and retains only the length-neutral quality signal. It causes the largest overall degradation in both metrics, showing that explicitly modeling the utility of additional generation is a major source of the improvement. (iii) *Dataset-Level Utility* replaces context-specific marginal utility with a single dataset-level utility signal, retaining an overall direction and strength for length utility while removing sample-level adaptation. Its consistent performance drop shows that dataset-level utility cannot fully capture how the value of additional generation varies across contexts. Overall, every ablation degrades performance on both benchmarks, supporting the combined use of length-neutral quality comparison and context-specific utility-guided length learning in UTIL.

6 Conclusion

Our analysis shows that additional response tokens do not have a universal effect on quality in LLM post-training: they can provide meaningful gains in some contexts, little return in others, and negative effects elsewhere. To capture this variation, we introduced marginal token utility and proposed UTIL, which uses the utility signal to adapt length-related learning in a context-dependent and bidirectional manner. Across models, training data settings, and alignment benchmarks, UTIL consistently improves alignment performance over representative length bias mitigation methods under a comparable overall generation amount, indicating that the benefit comes from using additional generation more selectively rather than simply encouraging shorter or longer responses. Overall, these results suggest that effective length control should account for both the cost of additional tokens and the quality gains they provide, treating response length as a context-dependent component of post-training rather than only a source of global bias.

AI USE STATEMENT

Generative AI tools were used to assist with writing and language refinement, literature discovery, and research execution, including feedback on experimental design, code development, and analysis. All AI-assisted content was reviewed and verified by the authors, including citations, code, and experimental results. The authors made the final decisions regarding the methodology, experiments, interpretation, and manuscript content, and take responsibility for the final work.

REPRODUCIBILITY STATEMENT

We provide the methodological details needed to reproduce UTIL in Section 4, including the definition and estimation of marginal token utility and its integration into preference optimization. Section 5 and Appendix A describe the datasets, training setup, baselines, evaluation benchmarks, and metrics used in our experiments. We use publicly available models, datasets, and evaluation benchmarks and report the corresponding model, data, and training configurations to facilitate reproduction of our results. Anonymous source code and experiment configurations are provided in the supplementary material.

REFERENCES

Anirudh Bharadwaj, Chaitanya Malaviya, Nitish Joshi, and Mark Yatskar. Flattery, fluff, and fog: Diagnosing and mitigating idiosyncratic biases in preference models. In *International Conference on Learning Representations*, volume 2026, pp. 19171–19191, 2026.

Yuyan Bu, Liangyu Huo, Yi Jing, and Qing Yang. Beyond excess and deficiency: Adaptive length bias mitigation in reward models for rlhf. In *Findings of the association for computational linguistics: NAACL 2025*, pp. 3091–3098, 2025.

Jianfeng Cai, Jinhua Zhu, Ruopei Sun, Yue Wang, Li Li, Wengang Zhou, and Houqiang Li. Disentangling length bias in preference learning via response-conditioned modeling. In *International Conference on Learning Representations*, volume 2026, pp. 93564–93598, 2026.

Changyu Chen, Zichen Liu, Chao Du, Tianyu Pang, Qian Liu, Arunesh Sinha, Pradeep Varakantham, and Min Lin. Bootstrapping language models with dpo implicit rewards. In *International Conference on Learning Representations*, volume 2025, pp. 55898–55915, 2025.

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 Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9722–9744. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. In *First Conference on Language Modeling*, 2024.

Benjamin Feuer, Micah Goldblum, Teresa Datta, Sanjana Nambiar, Raz Besaleli, Samuel Dooley, Max Cembalest, and John P Dickerson. Style outweighs substance: Failure modes of llm judges in alignment benchmarking. In *International Conference on Learning Representations*, volume 2025, pp. 11552–11582, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

Taneesh Gupta, Rahul Madhavan, Xuchao Zhang, Chetan Bansal, and Saravan Rajmohan. Refa: Reference free alignment with fine-grained length control. In *Second Conference on Language Modeling*, 2025.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. Explaining length bias in llm-based preference evaluations. In *EMNLP (Findings)*, pp. 6763–6794, 2025.

Zeyu Huang, Zihan Qiu, Edoardo M Ponti, Ivan Titov, et al. Post-hoc reward calibration: A case study on length bias. In *International Conference on Learning Representations*, volume 2025, pp. 37609–37635, 2025.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34209–34231. PMLR, 2025.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, et al. Rrm: Robust reward model training mitigates reward hacking. In *International Conference on Learning Representations*, volume 2025, pp. 62682–62700, 2025.

Junru Lu, Jiazheng Li, Siyu An, Meng Zhao, Yulan He, Di Yin, and Xing Sun. Eliminating biased length reliance of direct preference optimization via down-sampled kl divergence. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1047–1067, 2024.

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.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. A long way to go: Investigating length correlations in RLHF. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=G8LaO1P0xv.

Wenqian Ye, Guangtao Zheng, and Aidong Zhang. Rectifying shortcut behaviors in preference-based reward learning. *Advances in Neural Information Processing Systems*, 38:64712–64740, 2025.

Hee Suk Yoon, Eunseop Yoon, Mark Hasegawa-Johnson, Sungwoong Kim, and Chang D. Yoo. Confpo: exploiting policy model confidence for critical token selection in preference optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, ICML'25. JMLR.org, 2025.

Xuanchang Zhang, Wei Xiong, Lichang Chen, Tianyi Zhou, Heng Huang, and Tong Zhang. From lists to emojis: How format bias affects model alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26940–26961, 2025.

Kangwen Zhao, Jianfeng Cai, Jinhua Zhu, Ruopei Sun, Dongyun Xue, Wengang Zhou, Li Li, and Houqiang Li. Bias fitting to mitigate length bias of reward model in rlhf. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2912–2927, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A Implementation Details

Models and Training Data. We train meta-llama/Meta-Llama-3-8B-Instruct and Qwen/Qwen2.5-7B-Instruct under two complementary preference data settings. In the same-model setting, we use Llama-3 UltraFeedback-ArmoRM, where candidate responses are generated by the same Llama-3-8B-Instruct model used as the training backbone. This model-matched setting provides preference comparisons directly over responses from the target model, while requiring model-specific response data. We use the provided chosen–rejected pairs, giving 59,829 training pairs. In the cross-model setting, we use UltraFeedback, whose candidate responses are generated by multiple source models. This heterogeneous setting provides a reusable preference dataset that can be applied across different target models, although its response distribution is less closely matched to any individual backbone. We construct each pair by selecting the response with the highest overall score as preferred and a randomly sampled response with a different score as rejected, using a fixed seed, giving 63,269 training pairs. The same-model and cross-model settings contain five and four scored candidates per prompt, respectively. For UTIL, marginal token utility is estimated from all scored candidate responses available for each prompt, while retaining exactly the same number of preference pairs as the corresponding source dataset.

Training. DPO, R-DPO, SamPO, ConfPO, and UTIL are trained for one epoch with a learning rate of 5×10^{-6}, an effective batch size of 32, a maximum sequence length of 1,536 tokens, a maximum prompt length of 512 tokens, and random seed 42. We use $\beta = 0.1$ for all methods unless the baseline requires a different configuration; ConfPO uses $\beta = 1$ following its intended design requirement. R-DPO additionally introduces the length-regularization coefficient α. To ensure a fair comparison, we evaluate their recommended values $\alpha \in \{0.005, 0.01, 0.05\}$ and report the best-performing setting, $\alpha = 0.01$, in the main comparison. All training is performed on four NVIDIA L40S GPUs.

Generation and Evaluation. We use the same decoding configuration for all compared methods within each benchmark, with sampling temperature 0.7 and top-$p = 1.0$. AlpacaEval 2 generations allow up to 2,048 new tokens and are evaluated against the standard gpt4_1106_preview reference outputs, while Arena-Hard v0.1 generations allow up to 4,096 new tokens and are evaluated against the standard gpt-4-0314 outputs.

B Additional Results and Analysis

B.1 Statistical Analysis of Paired Improvements

Table 3 reports 95% paired-bootstrap confidence intervals for the win rate improvements of UTIL over DPO and SamPO, computed with 10,000 resamples. UTIL achieves clear positive gains in all eight comparisons, reaching up to $+4.30$ percentage points over DPO and $+3.60$ points over SamPO. Overall, the confidence intervals exclude zero in four of eight comparisons, including three of four comparisons against DPO, providing statistical support for the consistent win rate improvements observed in the main results. The wider intervals in several Qwen2.5-7B-Instruct cross-model comparisons are also consistent with the generally smaller differences among preference training methods in this setting, where the training responses are generated by heterogeneous models rather than the target model itself.

Table 3: Paired win-rate improvements of UTIL over the two strongest baselines, DPO and SamPO. We report changes in win rate (ΔWR) with 95% paired-bootstrap confidence intervals (CIs).

Training setting	Benchmark	UTIL − DPO	UTIL − SamPO
Llama-3-8B (Same-model data)	Arena-Hard	$+4.30$ [0.51, 8.22]	$+3.60$ [-0.41, 7.23]
	AlpacaEval 2	$+2.82$ [1.05, 4.61]	$+2.97$ [1.28, 4.66]
Qwen-2.5-7B (Cross-model data)	Arena-Hard	$+2.20$ [-1.21, 5.47]	$+2.50$ [-1.37, 5.32]
	AlpacaEval 2	$+3.33$ [1.29, 5.38]	$+0.33$ [-1.45, 2.09]

B.2 Qualitative Examples of Context-Dependent Token Utility

We further provide two qualitative examples to illustrate what positive and negative marginal token utility capture in individual contexts. Table 4 shows a positive-utility example in which additional generation contributes useful task-relevant coverage, while Table 5 shows a negative-utility example in which a shorter response is more focused on the requested comparison. Together, the examples illustrate why response length should be treated context-dependently rather than assigned a universal positive or negative effect.

Table 4: Qualitative example with *positive* marginal token utility. The longer preferred response uses additional generation to cover complicated conservation and accessibility challenges requested by the prompt.

	Shorter (less preferred)	**Longer (more preferred)**
Prompt	What efforts are being made to preserve and maintain Monticello, and what challenges does the site face in terms of conservation and accessibility?	
Response	"One of the main challenges faced by the Monticello site in terms of conservation and accessibility is the fact that it is located in a rural area, making it difficult to attract visitors and fundraise for maintenance. The foundation relies heavily on private donations and grants to fund the upkeep of the property. ...To address these challenges, the foundation has taken several steps to increase awareness and visitation of the property. They have implemented a variety of **marketing and outreach efforts** ...improved the visitor experience by upgrading facilities and offering educational programs and events. ...established partnerships with other organizations and institutions ..."	"The Monticello site faces several challenges in terms of conservation and accessibility, including: 1. **Weathering**: Monticello is exposed to the elements ...2. **Maintenance**: The historic estate requires constant maintenance and restoration ...3. **Preservation**: ...balance the need to preserve the past and the need to make it accessible to visitors. 4. **Accessibility**: The physical structure of Monticello can limit accessibility for those with disabilities. ...5. **Safety**: The site also faces challenges in terms of visitor safety ..."
Observation	Additional tokens are useful here. The prompt asks about multiple dimensions of conservation and accessibility. The **longer preferred response** uses its additional generation to introduce **several distinct and directly relevant challenges**, including weathering, maintenance, preservation, accessibility, and safety. In contrast, the **shorter response** mainly discusses funding, outreach, and visitor programs, providing **less complete coverage of the requested dimensions**. Thus, the additional tokens expand task-relevant coverage rather than merely repeating existing information.	

Table 5: Qualitative example with *negative* marginal token utility. The shorter preferred response concentrates its generation on the requested comparison, whereas the longer response devotes additional content to less central provider and service details.

	Longer (less preferred)	**Shorter (more preferred)**
Prompt	How does Debtor Finance work, and what are the main differences between traditional bank lending and Invoice Discounting?	
Response	"Business Debtor Finance, also known as Invoice Discounting, Factoring, or Debtor Finance, allows businesses to access immediate cash by leveraging their outstanding debts. . . . However, Debtor Finance is different from traditional bank lending in several aspects: . . . 3. **Fewer Restrictions**: Debtor Financing companies generally have fewer restrictions on the type of business they work with, including start-ups and businesses with tax arrears or no security to offer. 4. **Collaboration**: Debtor Financing companies work closely with you to ensure that new debtors and debtor limits are managed effectively. They may also provide assistance in the collection of larger debts, but they do not manage your business operations. 5. **Customizability**: Debtor Financing providers offer a range of options tailored to your business needs. They provide a free, obligation-free quote with up to four options suitable for your business. In summary, Debtor Financing is a powerful tool for businesses seeking to improve their cash flow, fuel growth, and maintain control over their operations. . . ."	"Debtor Finance or Invoice Discounting works by providing businesses with a cash advance on their outstanding invoices. This helps to improve the company's cash flow and enables growth by offering immediate access to capital. . . . The main differences between traditional bank lending and Invoice Discounting are: 2. **Speed**: Invoice Discounting provides access to funds within 24 hours of raising invoices, while traditional bank loans can take weeks or months to process and approve. 3. **Eligibility**: Invoice Discounting is open to new or old businesses and often accepts clients with tax arrears or no securities. Banks often require a business to have a minimum operating duration before being eligible for a loan, and they usually require collateral or property as security. 4. **Pricing**: Invoice Discounting tends to be more expensive than traditional bank lending . . . 5. **Credit Control**: Invoice Discounting allows businesses to retain their collections procedures, while banks may impose stricter requirements on credit controls and collection processes."
Observation	Additional tokens are less useful here. The prompt specifically asks for the differences between Invoice Discounting and traditional bank lending. The **shorter preferred response** uses its content to make **explicit side-by-side distinctions** in speed, eligibility, pricing, and credit control. In contrast, the **longer response** spends additional space on provider collaboration, customizability, and broader business benefits, which are informative but **less directly relevant to the requested comparison**. Thus, the additional generation increases response length without providing the same task-relevant quality gain.	