# MARS: Margin and Semantic-Aware Data Augmentation for Reward Modeling

Anonymous ACL submission


## Abstract

Reward modeling is central to alignment pipelines such as RLHF, RLAIF, and PPO-based policy optimization, yet its reliability is constrained by limited and heterogeneous human preference data that are expensive to collect at scale. While synthetic augmentation can expand preference supervision, existing methods often augment uniformly or at the representation level, without targeting examples where the reward model is uncertain or prone to mis-ranking. In this paper, we introduce MARS *(**M**argin and **S**emantic-**A**ware Data Augmentation for **R**eward Modeling)*, an adaptive augmentation framework that prioritizes low-margin preference pairs and uses semantic distance as a second layer for refinement to enhance the contrast between the chosen and rejected responses. Across multiple preference datasets, reward-model backbones, downstream alignment settings, and benchmarks including RewardBench and AlpacaEval, MARS improves both reward-model quality and alignment performance over existing baselines. Our results show that reward-model augmentation is most effective when guided by both model margins and semantic structure.


## 1 Introduction

The alignment of large language models (LLMs) has emerged as a central challenge as these models are increasingly deployed in several high-stakes domains such as education (Al Faraby et al., 2024; Alhafni et al., 2024), scientific research (Ren et al., 2025; Liao et al., 2024), healthcare (Yang et al., 2023; Cascella et al., 2023), and finance (Lakkaraju et al., 2023; Zhao et al., 2024). Contemporary alignment pipelines predominantly rely on human labeled preference data, where annotators provide pairwise comparisons over prompt-response tuples $(x, y^+, y^-)$, indicating a preferred response $y^+$ for input prompt $x$ over a rejected alternative $y^-$.

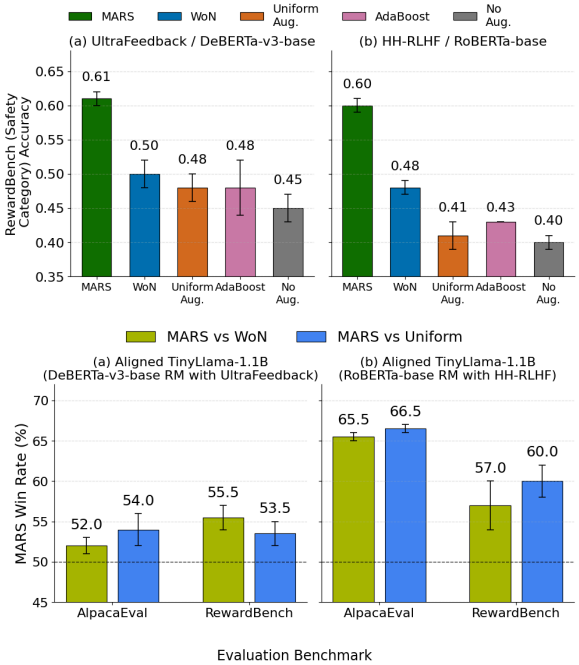


Figure 1: MARS improves RewardBench (Safety category) accuracy over reward-model baselines and yields stronger alignment win rates against WoN and Uniform on AlpacaEval and RewardBench. Error bars denote standard deviation across seeds (see Table 2 for details).

A large class of alignment methods is policy-based, including TRPO (Schulman et al., 2015) and PPO (Schulman et al., 2017), where a learned reward model guides policy updates. Reward-model-free alternatives such as DPO (Rafailov et al., 2023) bypass this by optimizing policies directly from preference data. Nonetheless, production-scale pipelines including RLHF (Ouyang et al., 2022) and RLAIF (Lee et al., 2023) continue to rely on PPO-style optimization, making reward model quality a decisive and key factor in shaping the aligned policy.

Recent studies have identified key vulnerabilities in reward modeling, reward hacking, misgeneralization, and sensitivity to spurious correlations (Shen et al., 2023), and empirical evidence shows that even minor perturbations, such as appending boil-

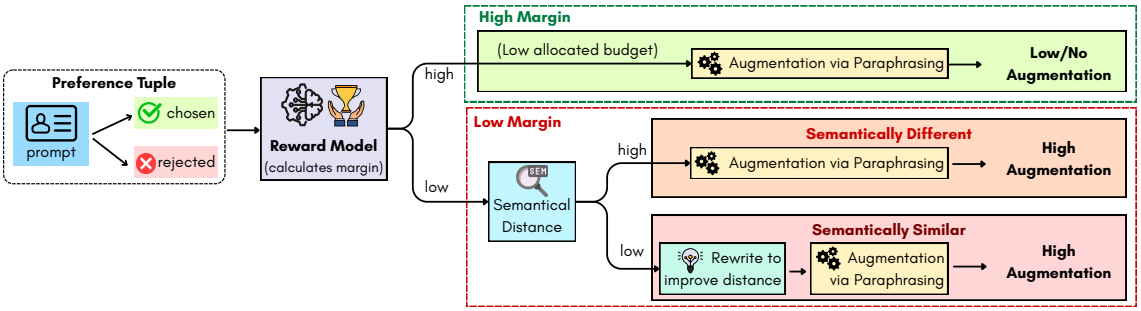


Figure 2: Overview of MARS: high-margin preference pairs receive limited (or no) augmentation, while low-margin pairs are further evaluated using semantic distance. Low-margin pairs with sufficient semantic separation receive high augmentation, whereas semantically close pairs are first rewritten to increase chosen-rejected separation before generating synthetic preference samples.

erplate text to rejected responses, can substantially alter reward predictions (Hughes et al., 2024).

Motivated by these limitations, prior work has explored robustness techniques such as contrastive sentence embedding consistency (Gao et al., 2021), clustering-based consistency across augmented views (Caron et al., 2020), and causal artifact mitigation for reward modeling (Liu et al., 2024a). To reduce reliance on costly human preference annotation, reward-based selection methods such as Best-of-$N$ (Yang et al., 2024; Gui et al., 2024) and West-of-$N$ (Pace et al., 2024) use generated candidates for policy improvement or reward-model self-training. In contrast, SimCSE (Gao et al., 2021) and SwAV (Caron et al., 2020) provide representation-level consistency rather than direct preference-pair selection for reward-model augmentation. Despite their empirical success, many existing augmentation strategies remain only indirectly tied to the reward model's failure modes. Although the final selection of synthetic pairs may use reward-model scores (Yang et al., 2024; Pace et al., 2024), the augmentation process itself is typically reward-model agnostic: synthetic variants are generated uniformly or from candidate pools before identifying where the current reward model is uncertain, or misranks responses. This motivates a model-aware augmentation strategy that directs supervision toward low-margin regions where the reward model is uncertain or prone to misranking.

Motivated by these observations, we propose MARS *(**M**argin and **S**emantic-**A**ware Data Augmentation for **R**eward Modeling)*, a self-refining framework that couples synthetic data generation with reward-model learning dynamics. MARS concentrates augmentation on low-margin or mis-ranked examples and uses semantic contrast to determine whether selected pairs should be directly augmented or first rewritten to preserve a clear chosen-rejected distinction (for detailed related work we refer the readers to Section 2 and Appendix A.1). The key contributions of this paper are as follows:

(1) ***Margin and semantic-aware augmentation.*** We present MARS, a data augmentation framework that allocates synthetic preference augmentation to low-margin or mis-ranked comparisons where the reward model is uncertain, while using semantic distance to ensure that augmented pairs remain informative.

(2) ***Interpreting margin-based allocation.*** We show that the MARS augmentation rule can be interpreted as the solution to a KL-regularized reweighting objective, supporting greater augmentation emphasis on low-margin pairs. This allocation is coupled with semantic-distance-aware refinement, which helps to sharpen the chosen-rejected distinction before augmentation.

(3) ***Empirical validation across reward modeling and alignment.*** Across multiple benchmarks, datasets and reward model backbones, MARS improves reward model quality. When used for downstream policy alignment, MARS-trained reward model-based aligned models also yield stronger win rates than the existing baselines. Experimental analysis (see Section 4) on AlpacaEval (Li et al., 2023) and RewardBench (Lambert et al., 2024) shows consistent gains of MARS (see Figure 1).

## 2 Related Work

In this section, we review prior work on preference-based reward modeling and augmentation strategies, positioning MARS within the broader landscape of adaptive reward-model training.

**Reward Modeling from Preferences.** Reward modeling is central to reward-based alignment

| Property / Selection Criterion | Uniform Aug. | AdaBoost | BoN | WoN | MARS |
|---|---|---|---|---|---|
| Operates during reward-model training | ✓ | ✓ | ✗ | ✓ | ✓ |
| Uses synthetic preference augmentation | ✓ | ✗ | ✗ | ✓ | ✓ |
| Selection driven by reward model scores | ✗ | ✗ | ✓ | ✓ | ✓ |
| Targets low-margin / uncertain comparisons | ✗ | Partial | ✗ | ✗ | ✓ |
| Uses margin-aware augmentation allocation | ✗ | ✗ | ✗ | ✗ | ✓ |
| Uses semantic-aware refinement | ✗ | ✗ | ✗ | ✗ | ✓ |
| Adaptive across training epochs | ✗ | ✓ | ✗ | Partial | ✓ |

Table 1: The table distinguishes whether a method operates during reward-model training, uses synthetic preference augmentation, and adaptively selects examples based on reward-model uncertainty. Unlike Uniform Augmentation, AdaBoost-style reweighting, BoN, and WoN, MARS first identifies low-margin preference pairs and then applies margin- and semantic-aware augmentation while preserving the original preference data.

pipelines such as TRPO (Schulman et al., 2015), PPO (Schulman et al., 2017), RLHF (Ouyang et al., 2022), and RLAIF (Lee et al., 2023). Given a prompt $x$ and two responses $(y^+, y^-)$, a parameterized reward model $r_\theta(x, y)$, with trainable parameters $\theta$, is trained to approximate the latent human preference function $r^*(x, y)$. Under the Bradley-Terry (BT) model (Bradley and Terry, 1952), and related ranking models such as Plackett-Luce (Plackett, 1975; Luce et al., 1959), the probability that $y^+$ is preferred over $y^-$ is

$$p(y^+ \succ y^- \mid x; \theta) = \sigma\big(r_\theta(x, y^+) - r_\theta(x, y^-)\big)\,,$$

where $\sigma(\cdot)$ denotes the logistic sigmoid function, and the reward model is trained by minimizing the negative log-likelihood over preference tuples $z = (x, y^+, y^-)$:

$$\mathcal{L}(\theta) = -\mathbb{E}_{z\sim\mathcal{D}}\big[\log \sigma\big(r_\theta(x, y^+) - r_\theta(x, y^-)\big)\big]\,.$$

**Augmentation for Reward Modeling.** Reward models trained on limited human preference data can exploit spurious correlations, while collecting diverse human annotations is costly (Gao et al., 2021; Caron et al., 2020; Liu et al., 2024a). This has motivated augmentation, selection, and self-training strategies for reward modeling. As summarized in Table 1, these methods differ mainly in *when* examples are selected and *what signal* drives that selection. Uniform augmentation expands all preference pairs equally, without targeting regions where the reward model is uncertain. Best-of-$N$ (Gui et al., 2024; Dong et al., 2023; Sessa et al., 2024) samples multiple candidate responses and selects high-reward outputs, primarily for inference-time selection, distillation, or policy improvement. West-of-$N$ (Pace et al., 2024) is more directly tied to reward-model self-training since it constructs synthetic preference pairs from high-confidence best-worst candidates and uses them to retrain the reward model. Thus, BoN and WoN both rely on reward-based selection over generated candidates, but they do not explicitly target existing low-margin preference pairs for augmentation. In contrast, MARS first uses reward margins to identify low-margin or mis-ranked preference pairs, then adaptively augments those. As shown in Table 1, it uniquely combines margin-aware allocation, and semantic-aware refinement while preserving the original preference pairs.

Other augmentation and robustness methods are complementary to this selection mechanism. SimCSE (Gao et al., 2021) and SwAV (Caron et al., 2020) provide representation-level consistency, while RRM (Liu et al., 2024a) improves reward-model robustness against prompt-independent artifacts; in contrast, MARS determines where synthetic preference augmentation should be concentrated. These methods primarily specify *how* to improve representations or reduce artifacts; MARS specifies *where* augmentation should be concentrated. AdaBoost-style training (Freund and Schapire, 1997) also emphasizes difficult examples by reweighting low-margin, but it does not generate augmented preference variants. MARS therefore builds on the broader intuition behind hard-example and importance-weighted training (Shrivastava et al., 2016; Katharopoulos and Fleuret, 2018), while incorporating semantic-distance-aware margins (Mohri and Zhong, 2026), uncertainty-based preference selection (Muldrew et al., 2024).

## 3 MARS: Margin and Semantic-Aware Data Augmentation for Reward Modeling

We now present MARS, our framework for adaptive preference augmentation in reward modeling. The framework consists of two major components:

1. **Margin-aware augmentation.** Given the current reward model, MARS computes per-sample reward margins to identify low-margin or mis-ranked preference pairs where the current reward model is uncertain or likely to fail. These hard examples are prioritized because they provide more informative corrective supervision than already well-separated pairs, aligning with the prior work that focus on uncertain or confidently incorrect comparisons (Muldrew et al., 2024). Under a fixed synthetic-data budget, MARS therefore allocates augmentation non-uniformly, concentrating more samples on low-margin regions.

2. **Semantic-aware refinement.** Margin alone does not determine whether a low-margin pair provides a useful training signal: if the chosen and rejected responses are semantically near identical, naive augmentation may simply reproduce weak or ambiguous comparisons. Motivated by the structure-aware preference learning principle that margins should account for semantic distance (Mohri and Zhong, 2026), MARS also measures the semantic distance between the chosen and rejected responses among selected low-margin pairs. Pairs with sufficient semantic separation are directly augmented, while semantically close pairs are first rewritten to sharpen the chosen-rejected contrast before augmentation.

### 3.1 Margin-Aware Augmentation Allocation

Let $\mathcal{D} = \{z_i\}_{i=1}^N$ denote a fixed human-labeled preference dataset, where each tuple $z_i = (x_i, y_i^+, y_i^-)$ consists of a prompt $x_i$, a chosen response $y_i^+$, and a rejected response $y_i^-$. We train a reward model $r_\theta$ parameterized by $\theta$ over $T$ epochs, adaptively refining the augmented training distribution based on the model's evolving behavior.

**Analyzing Reward Margin.** At epoch $t$, we compute the reward margin for each tuple $z_i$ as:

$$\Delta_i^t := r_\theta^t(x_i, y_i^+) - r_\theta^t(x_i, y_i^-). \tag{1}$$

A large positive margin indicates that the reward model confidently ranks $y_i^+$ over $y_i^-$; such pairs already provide a clear preference signal and therefore require limited additional augmentation. Importantly, MARS does not discard these original pairs, so the existing supervision is preserved throughout training. Conversely, a small or negative margin signals an ambiguous or mis-ranked pair, precisely the region where the model's decision boundary is unreliable and where concentrated synthetic supervision is most valuable.

**Adaptive Augmentation Budget $(B)$ Allocation.** We introduce an epoch-level augmentation budget $B^t$, capping the total number of synthetic samples generated from $\mathcal{D}$ at epoch $t$. Rather than distributing this budget uniformly, MARS allocates it proportionally to each tuple's difficulty. Formally, the augmentation probability for the $i^{\text{th}}$ tuple is defined via a softmax over the negated margins:

$$q_i^t = \frac{\exp(-\tau\, \Delta_i^t)}{\sum_j \exp(-\tau\, \Delta_j^t)}, \tag{2}$$

where $\tau > 0$ controls the sharpness of the allocation. Since $\sum_i q_i^t = 1$, the quantity $b_i^t = B^t q_i^t$ specifies the fraction of the epoch-level augmentation budget $B^t$ assigned to tuple $z_i$. Larger values of $\tau$ concentrate the budget more strongly on low-margin examples, whereas smaller values approach uniform allocation, recovering standard augmentation as a limiting case.

For each tuple $z_i$, the margin-derived budget $b_i^t$ specifies how much synthetic generation is allocated to that tuple at epoch $t$. In MARS, we split this budget between the chosen and rejected responses by selecting $n_i^+$ and $n_i^-$ such that

$$n_i^+ + n_i^- = b_i^t. \tag{3}$$

In MARS allocation, we set $n_i^+ = \lfloor b_i^t/2 \rfloor$ and $n_i^- = b_i^t - n_i^+$. These variants yield up to $(n_i^+ + 1)(n_i^- + 1)$ candidate preference pairs for the same prompt, including the original comparison. Thus, low-margin or mis-ranked tuples receive a larger local augmentation pool, while confidently separated tuples receive fewer generated variants.

We now show that the MARS allocation in Equation (2) is also an unique optimizer of a principled variational objective. Given preference dataset $\mathcal{D} = \{z_i\}_{i=1}^N$ with $N$-samples, for $i = 1, \ldots, N$, we define $P_N$ as the uniform empirical distribution over $\mathcal{D} : P_N(z_i) = \frac{1}{N}$. Thus, $P_N$ corresponds to the standard uniform augmentation baseline. Let $\Delta_N = \{Q \in \mathbb{R}^N : \sum_{i=1}^N Q(z_i) = 1\}$ denote the probability simplex over the training tuples, and let $\Delta_\theta(z)$ be the reward margin of tuple $z$ under the current reward model. We seek a reweighting distribution $Q \in \Delta_N$ that assigns more mass to low-margin examples while remaining close to $P_N$,

this boils down to this optimization problem:

$$\max_{Q \in \Delta_N} \left\{ -\mathbb{E}_{z\sim Q}\big[\Delta_\theta(z)\big] \;-\; \frac{1}{\tau} D_{\mathrm{KL}}(Q \parallel P_N) \right\}. \tag{4}$$

The first term encourages $Q$ to assign higher weight to low-margin examples, and the KL penalty prevents degenerate concentration by anchoring $Q$ near the uniform baseline. The parameter $\tau$ controls how strongly the allocation favors low-margin examples: larger values place more mass on harder tuples, while smaller values keep the distribution closer to the uniform empirical baseline $P_N$. This softmax form is consistent with standard KL-regularized variational objectives, where Gibbs or exponential-tilted distributions arise naturally (Ziebart et al., 2008; Haarnoja et al., 2017).

**Lemma 1** (MARS as KL-Regularized Reweighting)**.** *The optimization problem in* (4) *admits a unique closed-form solution, given $P_N$ is uniform over $\mathcal{D}$, this solution simplifies to:*

$$Q^*_\theta(z_i) \;=\; \frac{\exp\big(-\tau\,\Delta_i(\theta)\big)}{\sum_{j=1}^{N} \exp\big(-\tau\,\Delta_j(\theta)\big)}, \tag{5}$$

*which coincides exactly with the* MARS *augmentation allocation rule in Equation* (2)*.*

Lemma 1 gives a variational interpretation of the MARS allocation rules as an application of a known variational result, its value in this work is interpretive: it provides a principled justification for the MARS allocation rule, connects it to the broader family of hard-example and importance-weighted training strategies such as OHEM (Shrivastava et al., 2016) and exponential importance sampling (Katharopoulos and Fleuret, 2018). Full proof is provided in Appendix A.2.

### 3.2 Semantic-Aware Refinement

Margin-based allocation in Equation 2 determines *where* augmentation should be concentrated, but margin alone does not determine *how* augmentation should be performed. Two preference tuples may have similar reward margins while exhibiting very different semantic relationships between the chosen and rejected responses. If a low-margin pair already contains a clear semantic contrast, it represents a useful decision-boundary example and can be directly augmented. However, if the chosen and rejected responses are semantically near-identical, naive paraphrasing may simply reproduce the same ambiguity providing limited additional supervision.

To address this, MARS combines margin-based budgeting with semantic-distance-aware refinement. To determine whether a pair should be directly augmented or first refined, we compute the semantic distance between $y_i^+$ and $y_i^-$. Let $f(\cdot)$ denote a pretrained sentence-transformer encoder (such as `all-mpnet-base-v2`; (Reimers and Gurevych, 2019)), and let $\mathbf{e}_i^+ = f(y_i^+)$ and $\mathbf{e}_i^- = f(y_i^-)$ be unit-normalized response embeddings. We define the inner product and distance as:

$$s_i = \mathbf{e}_i^+ \cdot \mathbf{e}_i^-, \qquad d_i = 1 - s_i, \tag{6}$$

where a larger distance $d_i$ indicates greater semantic separation between the chosen and rejected responses (detailed analysis is presented in Appendix A.3). We then use the dataset-level mean distance $\bar{d} = \frac{1}{N}\sum_{i=1}^{N} d_i$ as a surrogate threshold and assign each tuple a binary semantic distance label:

$$\ell_i = \begin{cases} \texttt{high}, & d_i \geq \bar{d}, \\ \texttt{low}, & d_i < \bar{d}. \end{cases} \tag{7}$$

For tuples with $\ell_i = \texttt{high}$, the chosen and rejected responses are already sufficiently distinct. Therefore, MARS directly paraphrases these responses to create diverse synthetic variants while preserving the original preference contrast. Tuples with $\ell_i = \texttt{low}$ are first rewritten using an external model (`GPT 4.1` (OpenAI, 2025)) to *increase semantic separation between the chosen and rejected responses* while preserving the original preference label. Since rewriting can inadvertently change the intended comparison, MARS retains only rewritten pairs that preserve the chosen-rejected preference relation and do not introduce harmful or undesired content (see Appendix A.3 for details). The full MARS framework, integrating margin computation, semantic filtering, response rewriting, and augmented retraining, and reward modeling is summarized in Algorithm 1.

**Remark.** Although implemented with rewriting and paraphrasing, MARS is augmentation strategy-agnostic and can incorporate any representation-level perturbations (eg. (Gao et al., 2021), clustering-based consistency regularization (Caron et al., 2020)). Thus, it decouples *where* to augment from *how* to augment, forming a general framework for reward modeling.

**Algorithm 1:** MARS: Margin and Semantic-Aware Reward Modeling via Self-Refinement

**Input:** Preference dataset $\mathcal{D} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^N$, number of epochs $T$, reward model $r_\theta^t$, temperature $\tau$
**Output:** Dataset for the $t^{\text{th}}$ epoch, $\mathcal{D}^t$

**Initialize**: $\mathcal{D}^0 = \mathcal{D}$ as the human-labeled preference dataset, and initialize the reward model with an off-the-shelf model $r_\theta^0$;
**for** $t = 1$ **to** $T$ **do**
  **for** ***each*** $i^{th}$ *tuple* $z_i = (x_i, y_i^+, y_i^-)$ *from* $\mathcal{D}^{t-1}$, *where* $i = 1, 2, \ldots, N$ **do**
    Calculate chosen reward: $r_\theta^{t-1}(x_i, y_i^+)$;
    Calculate rejected reward: $r_\theta^{t-1}(x_i, y_i^-)$;
    Calculate reward margin: $\Delta_i^t = r_\theta^{t-1}(x_i, y_i^+) - r_\theta^{t-1}(x_i, y_i^-)$;
    Compute margin-aware sampling probability: $q_i^t = \frac{\exp(-\tau \Delta_i^t)}{\sum_j \exp(-\tau \Delta_j^t)}$;
    Assign augmentation budget $b_i^t = B^t q_i^t$ for tuple $z_i$ such that $n_i^+ + n_i^- = b_i^t$;
    **if** $d(y_i^+, y_i^-) \geq \frac{1}{N} \sum_{j=1}^N d_j$ **then**
      Generate $n_i^+$ paraphrases of $y_i^+$ and $n_i^-$ paraphrases of $y_i^-$;
    **else**
      Refine the low-distance pair to sharpen chosen-rejected separation while preserving the preference label and filtering harmful content;
      Generate $n_i^+$ paraphrases of the rewritten chosen response ($y_i^+$) and $n_i^-$ paraphrases of the rewritten rejected response ($y_i^-$);
    **end**
    Update $\mathcal{D}_{\text{syn}}$ by adding $b_i^t$ complete synthetic preference pairs sampled from the augmentation pool associated with $z_i$, and set $\mathcal{D}^t \leftarrow \mathcal{D}^{t-1} \cup \mathcal{D}_{\text{syn}}$;
  **end**
  Train reward model $r_\theta^{t-1}$ on dataset $\mathcal{D}^t$ to obtain $r_\theta^t$;
**end**
**return** $r_\theta^T$

# 4 Experimental Evaluation

In this section, we present a comprehensive empirical validation of the proposed framework MARS via reward model and alignment evaluation. For additional experimental details, and ablation study results we refer the readers to Appendix A.3.

We evaluate MARS along two dimensions: reward model quality and downstream policy alignment. Specifically, our experiments examine: (1) whether MARS improves reward modeling over existing augmentation and self-training baselines, (2) the contribution of semantic-aware refinement beyond margin-aware augmentation alone, (3) the effect of reward-model backbone size, and (4) whether improvements in reward modeling translate to stronger downstream alignment performance.

**Experimental Setup.** For reward modeling, we have trained reward-model-DeBERTa-v3-base and RoBERTa-base models on `HH-RLHF` (Bai et al., 2022), `UltraFeedback` (Cui et al., 2023), and `PKU-SafeRLHF` (Ji et al., 2024), comparing against 4 strategies: (a) Uniform Augmentation, (b) West-of-$N$ (WoN), (c) AdaBoost-style reweighting, and (d) No Augmentation on RewardBench (Lambert et al., 2024) benchmark. For downstream alignment, we use the trained reward models to guide PPO-style optimization of `Llama-3.2-1B` (Liu et al., 2024b) and `TinyLlama-1.1B`, and evaluate the aligned policies on AlpacaEval (Li et al., 2023) and RewardBench (Lambert et al., 2024) with `GPT-4.1` (OpenAI, 2025) as the pairwise judge. Synthetic augmented preference variants are generated using a T5-base `chatgpt-paraphraser`. In MARS, semantically close low-margin pairs are rewritten before augmentation using `GPT 4.1` (OpenAI, 2025) model. All methods use the same training and evaluation budgets, and results are reported as mean $\pm$ standard deviation across seeds. Code is available at this link. For detailed experimental setup, we refer the readers to Appendix A.3.

## 4.1 Improved Reward Modeling

Table 2 reports RewardBench (Lambert et al., 2024) results across three training datasets and two reward model backbones. Each category score measures the fraction of preference examples for which the reward model assigns a higher score to the human-preferred response than to the rejected response; thus, higher values indicate better pairwise preference accuracy. The reported Average is the average score over the Chat, ChatHard, Safety, and Reasoning categories. MARS *achieves the highest average score* in all backbone-dataset settings, indicating consistent improvements over Uniform Augmentation, WoN, AdaBoost, and No Augmentation. The gains are most pronounced on harder alignment-relevant categories such as ChatHard, Safety, and Reasoning, while some baselines occasionally obtain higher Chat scores. This pattern suggests that MARS does not merely improve performance on easier preference distinctions; rather, margin and semantic-aware augmentation provides a more balanced reward signal across evaluation dimensions.

| RM Training Dataset | Method | Chat | ChatHard | Safety | Reasoning | Average |
|---|---|---|---|---|---|---|
| | | | **DeBERTa-v3-base** | | | |
| HH-RLHF | MARS | $0.67 \pm 0.03$ | $\mathbf{0.55 \pm 0.04}$ | $0.56 \pm 0.03$ | $0.46 \pm 0.01$ | $\mathbf{0.56 \pm 0.02}$ |
| | Uniform Aug. | $0.70 \pm 0.01$ | $0.35 \pm 0.17$ | $0.57 \pm 0.03$ | $\mathbf{0.49 \pm 0.01}$ | $0.53 \pm 0.05$ |
| | No Aug. | $\mathbf{0.73 \pm 0.01}$ | $0.37 \pm 0.04$ | $0.58 \pm 0.05$ | $0.45 \pm 0.01$ | $0.53 \pm 0.02$ |
| | WoN | $0.55 \pm 0.01$ | $0.40 \pm 0.00$ | $\mathbf{0.58 \pm 0.02}$ | $0.47 \pm 0.01$ | $0.50 \pm 0.01$ |
| | AdaBoost | $0.72 \pm 0.02$ | $0.40 \pm 0.04$ | $0.52 \pm 0.01$ | $0.48 \pm 0.02$ | $0.53 \pm 0.02$ |
| PKU-SafeRLHF | MARS | $0.79 \pm 0.01$ | $\mathbf{0.46 \pm 0.02}$ | $\mathbf{0.64 \pm 0.01}$ | $0.49 \pm 0.03$ | $\mathbf{0.59 \pm 0.01}$ |
| | Uniform Aug. | $0.81 \pm 0.01$ | $0.31 \pm 0.00$ | $0.63 \pm 0.01$ | $\mathbf{0.54 \pm 0.03}$ | $0.57 \pm 0.01$ |
| | No Aug. | $\mathbf{0.85 \pm 0.01}$ | $0.31 \pm 0.02$ | $0.59 \pm 0.02$ | $0.49 \pm 0.01$ | $0.56 \pm 0.01$ |
| | WoN | $0.81 \pm 0.01$ | $0.34 \pm 0.05$ | $0.59 \pm 0.00$ | $0.49 \pm 0.02$ | $0.56 \pm 0.01$ |
| | AdaBoost | $0.84 \pm 0.00$ | $0.25 \pm 0.03$ | $0.63 \pm 0.02$ | $0.48 \pm 0.05$ | $0.55 \pm 0.01$ |
| UltraFeedback | MARS | $0.76 \pm 0.01$ | $\mathbf{0.47 \pm 0.01}$ | $\mathbf{0.61 \pm 0.01}$ | $0.52 \pm 0.03$ | $\mathbf{0.59 \pm 0.01}$ |
| | Uniform Aug. | $0.83 \pm 0.01$ | $0.41 \pm 0.01$ | $0.48 \pm 0.02$ | $0.50 \pm 0.11$ | $0.56 \pm 0.02$ |
| | No Aug. | $0.83 \pm 0.01$ | $0.36 \pm 0.05$ | $0.45 \pm 0.02$ | $\mathbf{0.54 \pm 0.06}$ | $0.55 \pm 0.01$ |
| | WoN | $0.83 \pm 0.01$ | $0.36 \pm 0.05$ | $0.50 \pm 0.02$ | $0.50 \pm 0.05$ | $0.55 \pm 0.01$ |
| | AdaBoost | $\mathbf{0.89 \pm 0.01}$ | $0.33 \pm 0.02$ | $0.48 \pm 0.04$ | $0.51 \pm 0.07$ | $0.55 \pm 0.00$ |
| | | | **RoBERTa-base** | | | |
| HH-RLHF | MARS | $0.50 \pm 0.02$ | $0.42 \pm 0.04$ | $\mathbf{0.60 \pm 0.01}$ | $\mathbf{0.56 \pm 0.05}$ | $\mathbf{0.52 \pm 0.03}$ |
| | Uniform Aug. | $0.50 \pm 0.01$ | $0.48 \pm 0.02$ | $0.41 \pm 0.02$ | $0.49 \pm 0.02$ | $0.47 \pm 0.01$ |
| | No Aug. | $\mathbf{0.52 \pm 0.01}$ | $0.47 \pm 0.12$ | $0.40 \pm 0.01$ | $0.48 \pm 0.01$ | $0.47 \pm 0.03$ |
| | WoN | $0.39 \pm 0.01$ | $\mathbf{0.54 \pm 0.07}$ | $0.48 \pm 0.01$ | $0.45 \pm 0.07$ | $0.46 \pm 0.03$ |
| | AdaBoost | $0.48 \pm 0.02$ | $0.40 \pm 0.04$ | $0.43 \pm 0.00$ | $0.49 \pm 0.00$ | $0.45 \pm 0.01$ |
| PKU-SafeRLHF | MARS | $0.58 \pm 0.01$ | $\mathbf{0.51 \pm 0.00}$ | $0.66 \pm 0.01$ | $\mathbf{0.49 \pm 0.01}$ | $\mathbf{0.56 \pm 0.01}$ |
| | Uniform Aug. | $0.65 \pm 0.00$ | $0.43 \pm 0.05$ | $\mathbf{0.68 \pm 0.02}$ | $0.35 \pm 0.05$ | $0.53 \pm 0.01$ |
| | No Aug. | $0.74 \pm 0.01$ | $0.31 \pm 0.00$ | $0.54 \pm 0.01$ | $0.46 \pm 0.01$ | $0.51 \pm 0.01$ |
| | WoN | $0.66 \pm 0.01$ | $0.43 \pm 0.05$ | $0.65 \pm 0.03$ | $0.41 \pm 0.03$ | $0.54 \pm 0.03$ |
| | AdaBoost | $\mathbf{0.76 \pm 0.01}$ | $0.37 \pm 0.00$ | $0.57 \pm 0.01$ | $0.39 \pm 0.01$ | $0.52 \pm 0.00$ |
| UltraFeedback | MARS | $0.75 \pm 0.01$ | $\mathbf{0.45 \pm 0.05}$ | $\mathbf{0.55 \pm 0.04}$ | $\mathbf{0.57 \pm 0.01}$ | $\mathbf{0.58 \pm 0.01}$ |
| | Uniform Aug. | $0.82 \pm 0.03$ | $0.38 \pm 0.01$ | $0.52 \pm 0.01$ | $0.50 \pm 0.01$ | $0.56 \pm 0.01$ |
| | No Aug. | $\mathbf{0.85 \pm 0.01}$ | $0.43 \pm 0.01$ | $0.33 \pm 0.02$ | $0.40 \pm 0.01$ | $0.50 \pm 0.01$ |
| | WoN | $0.79 \pm 0.01$ | $0.44 \pm 0.02$ | $0.39 \pm 0.01$ | $0.38 \pm 0.02$ | $0.50 \pm 0.01$ |
| | AdaBoost | $0.69 \pm 0.01$ | $0.37 \pm 0.02$ | $0.28 \pm 0.01$ | $0.48 \pm 0.01$ | $0.46 \pm 0.01$ |

Table 2: RewardBench-based comparison of MARS (this paper), West-of-$N$ (WoN), Uniform Augmentation, AdaBoost, and No Augmentation across two reward model backbones and three training datasets. Results are reported as mean $\pm$ standard deviation across different seeds under the same evaluation hyperparameters. Bold indicates the best score within each backbone, dataset group and metric.

### 4.2 Impact of Semantic-Aware Refinement

Figure 3(a) isolates the role of semantic-aware refinement. Across both model backbones, MARS consistently matches or improves over the variant without semantic refinement. This suggests that selecting low-margin examples is useful, but not sufficient: the selected pairs must also preserve a clear semantic contrast between the chosen and rejected responses. When a low-margin pair is semantically close, naive paraphrasing can produce redundant variants that retain the same ambiguity and provide weak ranking supervision. MARS addresses this by rewriting such pairs before augmentation, sharpening the chosen-rejected distinction and producing more informative synthetic preference pairs. Detailed results are reported in Appendix A.3.

### 4.3 Benefits for smaller RM backbones

Figure 3(b) suggests that MARS is very useful when reward-model (RM) backbone capacity is limited. Across the evaluated datasets, RoBERTa-base achieves comparable or stronger RewardBench average accuracy than RoBERTa-large under the MARS framework, indicating that improved reward modeling does not necessarily require scaling the reward backbone. Instead, MARS improves the effectiveness of the available supervision by concentrating augmentation on low-margin examples where the model needs the most corrective signal. This makes MARS attractive for resource-constrained alignment settings, where training, storing, or deploying larger reward models may be computationally expensive.

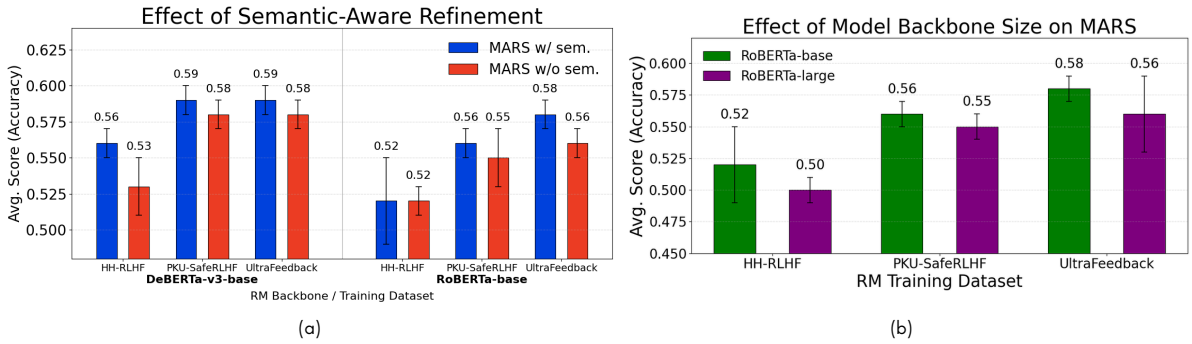


Figure 3: RewardBench average-score (classification accuracy of selecting chosen response over rejected) under two settings: (a) Effect of semantic-aware refinement, comparing full MARS with MARS without semantic analysis. (b) Effect of reward-model backbone size, comparing MARS with RoBERTa-base and RoBERTa-large backbones.

| | | **AlpacaEval** | | **RewardBench** | |
|---|---|---|---|---|---|
| **RM Backbone** | **Aligned Model** | MARS vs WoN | MARS vs Uniform | MARS vs WoN | MARS vs Uniform |
| **RM Training Dataset: HH-RLHF** | | | | | |
| DeBERTa | TinyLlama-1.1B | $(\mathbf{56.5}:43.5)\pm1.5$ | $(\mathbf{52.5}:47.5)\pm2.5$ | $(\mathbf{51.5}:48.5)\pm1.5$ | $(\mathbf{51.5}:48.5)\pm0.5$ |
| | Llama-3.2-1B | $(\mathbf{52.0}:48.0)\pm2.0$ | $(\mathbf{57.5}:42.5)\pm0.5$ | $(49.5:\mathbf{50.5})\pm1.5$ | $(\mathbf{57.0}:43.0)\pm1.0$ |
| RoBERTa | TinyLlama-1.1B | $(\mathbf{65.5}:34.5)\pm0.5$ | $(\mathbf{66.5}:33.5)\pm0.5$ | $(\mathbf{57.0}:43.0)\pm3.0$ | $(\mathbf{60.0}:40.0)\pm2.0$ |
| | Llama-3.2-1B | $(\mathbf{52.5}:47.5)\pm0.5$ | $(\mathbf{53.5}:46.5)\pm1.5$ | $(\mathbf{54.5}:45.5)\pm2.5$ | $(\mathbf{56.0}:44.0)\pm3.0$ |
| **RM Training Dataset: UltraFeedback** | | | | | |
| DeBERTa | TinyLlama-1.1B | $(\mathbf{52.0}:48.0)\pm1.0$ | $(\mathbf{54.0}:46.0)\pm2.0$ | $(\mathbf{55.5}:44.5)\pm1.5$ | $(\mathbf{53.5}:46.5)\pm1.5$ |
| | Llama-3.2-1B | $(\mathbf{51.0}:49.0)\pm1.0$ | $(\mathbf{52.0}:48.0)\pm1.0$ | $(\mathbf{53.0}:47.0)\pm1.0$ | $(\mathbf{56.5}:43.5)\pm1.5$ |
| RoBERTa | TinyLlama-1.1B | $(\mathbf{55.5}:44.5)\pm0.5$ | $(\mathbf{53.5}:46.5)\pm1.5$ | $(\mathbf{52.5}:47.5)\pm4.5$ | $(\mathbf{50.0}:50.0)\pm0.0$ |
| | Llama-3.2-1B | $(\mathbf{58.0}:42.0)\pm2.0$ | $(\mathbf{66.0}:34.0)\pm1.0$ | $(\mathbf{56.0}:44.0)\pm1.0$ | $(\mathbf{71.0}:29.0)\pm3.0$ |

Table 3: Downstream alignment comparison of MARS against WoN (Pace et al., 2024) and Uniform Augmentation across AlpacaEval (Li et al., 2023) and RewardBench (Lambert et al., 2024) using two RM backbones reward-model-DeBERTa-v3-base and RoBERTa-base. Results are reported as pairwise win rates for aligned models trained using reward models from HH-RLHF and UltraFeedback datasets.

### 4.4 Improved Downstream Model Alignment

Table 3 evaluates whether the reward-model gains from MARS translate into stronger downstream policy alignment. We compare policies optimized with MARS-trained reward models against policies optimized with reward models trained using WoN and Uniform Augmentation, reporting pairwise win rates on AlpacaEval (Li et al., 2023) and RewardBench (Lambert et al., 2024). MARS-aligned policies achieve consistently stronger win rates across reward-model backbones and policy models. The gains are especially notable because the same policy-optimization procedure is used across methods; the primary difference is the reward model used to guide alignment. This suggests that margin- and semantic-aware augmentation improves both reward-model accuracy and downstream alignment quality.

## 5 Conclusion

In this paper, we introduced MARS, a margin and semantic-aware framework for reward modeling. MARS targets augmentation toward uncertain or misranked preference pairs using reward-model margins, while semantic refinement preserves informative chosen-rejected contrast. Across datasets, reward-model backbones, and downstream alignment settings, MARS improves reward-model performance and yields stronger alignment outcomes on RewardBench and AlpacaEval compared to the existing baselines. These results suggest that effective reward-model augmentation depends not only on generating more preference data, but on placing supervision where the RM is most uncertain. Future work will extend this direction toward broader adaptive data selection and augmentation strategies for scalable applications.

## 6 Limitations

While MARS improves reward-model augmentation over existing baselines, it has some limitations as well. First, the margin-based allocation relies on reward estimates from the current checkpoint; early-training margins may be noisy before the model has calibrated, though this is naturally mitigated by the iterative refinement process. Second, the semantic refinement step assumes the paraphrasing model preserves preference labels while increasing chosen–rejected separation; we apply semantic filtering (Section 3.2) to guard against label-distorting rewrites, and leave more sophisticated verification mechanisms to future work. Third, our downstream alignment results rely on automated judges, a limitation shared across the alignment evaluation literature; human evaluation and broader judge sensitivity analyses remain important directions. Extending MARS to larger reward model and policy architectures, and evaluating on additional benchmarks, are natural next steps.

## 7 Ethical Statement

This study was conducted in compliance with relevant ethical guidelines and did not involve procedures requiring institutional ethical approval. The work uses publicly available preference datasets and synthetic augmentation methods for reward-model training. Since MARS generates rewritten and paraphrased preference responses, there is a potential risk of introducing mislabeled, biased, or harmful synthetic content if the generation process is not carefully filtered. To mitigate this, the semantic-refinement step (Section 3.2) is designed to preserve the original preference label and avoid harmful-intent amplification. More broadly, improved reward modeling can support safer and more reliable alignment pipelines, but it may also inherit biases from the underlying preference data, reward model backbones, and automatic evaluators. Future deployments should therefore include dataset auditing, safety filtering, and human oversight. We did not attempt to deanonymize any examples or link responses to real individuals, and any public release of augmented data should undergo automated PII screening, toxicity/safety filtering, and manual auditing.

**Potential Risks.** Because MARS synthetically generates paraphrased and rewritten preference responses, it may introduce label noise, biased content, or harmful synthetic examples if the rewriting process changes the intended preference relation. We attemp to mitigate this risk through semantic filtering and by retaining only rewritten pairs that preserve the original chosen-rejected preference label and avoid harmful-intent amplification. More broadly, the resulting reward models may inherit biases from the underlying preference datasets and automated evaluators, so deployment should include dataset auditing, safety filtering, and human oversight.

## 8 Information About Use of AI Assistants

We used AI assistants, including ChatGPT, Claude, during the preparation of this work. The use was limited to improving (1) the clarity, grammar, vocabulary and organization of the manuscript, as well as (2) assisting with debugging and refactoring portions of the experimental codebase. AI assistants were not used to generate experimental results, make scientific claims, or replace the authors' analysis and interpretation. All technical content, theoretical formulation, experimental design, reported results, and conclusions were developed, verified, and approved by the authors.

## 9 Artifact Licenses and Terms of Use.

We use publicly available datasets, models, and benchmarks, including HH-RLHF, UltraFeedback, PKU-SafeRLHF, RewardBench, AlpacaEval, sentence-transformer encoders, the T5-base paraphraser, and Hugging Face model checkpoints. These artifacts are used for research purposes and cited in Section 4. We do not redistribute the original datasets or third-party model weights; any released code provides scripts for preprocessing, augmentation, training, and evaluation. Users should verify and comply with the licenses and terms of use of each underlying dataset, model, benchmark, and API. The released code and generated artifacts are intended for research use in reward modeling, preference augmentation, and alignment evaluation, and should not be used outside the access conditions of the underlying artifacts.

## References


Said Al Faraby, Ade Romadhony, and 1 others. 2024. Analysis of llms for educational question classification and generation. *Computers and Education: Artificial Intelligence*, 7:100298.

Bashar Alhafni, Sowmya Vajjala, Stefano Bannò, Kaushal Kumar Maurya, and Ekaterina Kochmar. 2024. Llms in education: Novel perspectives, challenges, and opportunities. *arXiv preprint arXiv:2409.11917*.

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

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924.

Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. 2023. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of medical systems*, 47(1):33.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback, 2024. In *URL https://openreview. net/forum*.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Lin Gui, Cristina Gârbacea, and Victor Veitch. 2024. Bonbon alignment for large language models and the sweetness of best-of-n sampling. *Advances in Neural Information Processing Systems*, 37:2851–2885.

Yuan Guo, Peng Tian, Jayashree Kalpathy-Cramer, Susan Ostmo, J Peter Campbell, Michael F Chiang, Deniz Erdogmus, Jennifer G Dy, and Stratis Ioannidis. 2018. Experimental design under the bradley-terry model. In *IJCAI*, pages 2198–2204.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. 2017. Reinforcement learning with deep energy-based policies. In *Proceedings of the International Conference on Machine Learning*, pages 1352–1361.

John Hughes, Sara Price, Aengus Lynch, Rylan Schaeffer, Fazl Barez, Sanmi Koyejo, Henry Sleight, Erik Jones, Ethan Perez, and Mrinank Sharma. 2024. Best-of-n jailbreaking. *arXiv preprint arXiv:2412.03556*.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. 2024. Pku-saferlhf: A safety alignment preference dataset for llama family models. *arXiv e-prints*, pages arXiv–2406.

Angelos Katharopoulos and François Fleuret. 2018. Not all samples are created equal: Deep learning with importance sampling. In *International conference on machine learning*, pages 2525–2534. PMLR.

Kausik Lakkaraju, Sara E Jones, Sai Krishna Revanth Vuruma, Vishal Pallagani, Bharath C Muppasani, and Biplav Srivastava. 2023. Llms for financial advisement: A fairness and efficacy study in personal decision making. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 100–107.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Rewardbench: Evaluating reward models for language modeling. https://huggingface.co/spaces/allenai/reward-bench.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X Zhang. 2024. Llms as research tools: A large scale survey of researchers' usage and perceptions. *arXiv preprint arXiv:2411.05025*.

Tianqi Liu, Wei Xiong, Jie Ren, Lichang Chen, Junru Wu, Rishabh Joshi, Yang Gao, Jiaming Shen, Zhen Qin, Tianhe Yu, and 1 others. 2024a. Rrm: Robust reward model training mitigates reward hacking. *arXiv preprint arXiv:2409.13156*.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. 2024b. Spinquant: Llm quantization with learned rotations. *arXiv preprint arXiv:2405.16406*.

R Duncan Luce and 1 others. 1959. *Individual choice behavior*, volume 4. Wiley New York.

Mehryar Mohri and Yutao Zhong. 2026. Mind the gap: Structure-aware consistency in preference learning. *arXiv preprint arXiv:2604.27733*.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. 2024. Active preference learning for large language models. *arXiv preprint arXiv:2402.08114*.

OpenAI. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/. Accessed: 2025.

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

Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2024. West-of-n: Synthetic preferences for self-improving reward models. *arXiv e-prints*, pages arXiv–2401.

Robin L Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP*.

Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. 2025. Towards scientific intelligence: A survey of llm-based scientific agents. *arXiv preprint arXiv:2503.24047*.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Pier Giuseppe Sessa, Robert Dadashi, Léonard Hussenot, Johan Ferret, Nino Vieillard, Alexandre Ramé, Bobak Shariari, Sarah Perrin, Abe Friesen, Geoffrey Cideron, and 1 others. 2024. Bond: Aligning llms with best-of-n distillation. *arXiv preprint arXiv:2407.14622*.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large language model alignment: A survey. *arXiv preprint arXiv:2309.15025*.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769.

Joy Qiping Yang, Salman Salamatian, Ziteng Sun, Ananda Theertha Suresh, and Ahmad Beirami. 2024. Asymptotics of language model alignment. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 2027–2032. IEEE.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Yueqin Yin, Zhendong Wang, Yi Gu, Hai Huang, Weizhu Chen, and Mingyuan Zhou. 2024. Relative preference optimization: Enhancing llm alignment through contrasting responses across identical and diverse prompts. *arXiv preprint arXiv:2402.10958*.

Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Hanqi Jiang, and 1 others. 2024. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*.

Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. 2008. Maximum entropy inverse reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

# A APPENDIX

The Appendix for this paper is organised as follows:

A.1 Background, Motivation and Related Works
A.2 Proof of Lemma 1
A.3 Additional Experimental Results

## A.1 Additional Background and Related Work

This section provides additional context for the reward-modeling and augmentation baselines used in our experiments. The main paper discusses the high-level motivation for MARS; here, we focus on the technical relationship between preference-based reward modeling, synthetic preference construction, and hard-example weighting.

**Preference-based reward modeling.** Reward-based alignment pipelines such as RLHF and RLAIF commonly train a reward model from pairwise human or AI preference data and then use this model to guide policy optimization. Given a prompt $x$ and two candidate responses $(y^+, y^-)$, the reward model $r_\theta(x, y)$ is trained to assign a higher score to the preferred response. Under the Bradley–Terry model, the preference probability is modeled as

$$p(y^+ \succ y^- \mid x; \theta) = \sigma\big(r_\theta(x, y^+) - r_\theta(x, y^-)\big)\,, \tag{8}$$

and the model is trained by minimizing the corresponding negative log-likelihood over preference tuples. This formulation makes the reward margin

$$\Delta_\theta(z_i) = r_\theta(x_i, y_i^+) - r_\theta(x_i, y_i^-) \tag{9}$$

a natural diagnostic for reward-model confidence: large positive margins indicate confident preference separation, while small or negative margins indicate ambiguous or mis-ranked comparisons.

**Synthetic preference construction and selection.** Several recent approaches use synthetic data or reward-based selection to reduce reliance on costly human annotations. Uniform augmentation expands all preference pairs equally and therefore does not distinguish between already well-separated pairs and ambiguous examples. Best-of-$N$ methods sample multiple candidate responses and select high-reward outputs, typically for inference-time selection, policy improvement, or distillation. West-of-$N$ is more directly connected to reward-model self-training: it constructs synthetic preference pairs from high-confidence best–worst candidates and uses these pairs to further train the reward model. These methods rely on reward-based selection over generated candidates, but they do not explicitly allocate augmentation to existing preference pairs where the current reward model has low margin or makes an error.

**Representation-level robustness and complementary augmentation.** Representation-level methods such as SimCSE (Gao et al., 2021) and SwAV (Caron et al., 2020) encourage consistency across augmented views, while RRM (Liu et al., 2024a) improves reward-model robustness by mitigating prompt-independent artifacts. These approaches address how representations or reward models can be regularized, but they are not designed to decide where synthetic preference augmentation should be concentrated. In this sense, they are complementary to MARS: once low-margin pairs are selected, alternative augmentation operators such as representation-level perturbations, clustering-based consistency regularization, or artifact-aware filtering could be incorporated into the same margin-aware refinement loop.

**AdaBoost-style hard example weighting.** AdaBoost-style training provides a non-generative hard-example baseline. Instead of creating synthetic preference variants, it reweights the training loss to emphasize low-margin or mis-ranked pairs, for example using weights of the form

$$w_i^t \propto \exp\big(-\beta \Delta_\theta^t(z_i)\big)\,, \tag{10}$$

where $\beta$ controls the concentration on difficult examples. This is related to the margin-aware component of MARS, but differs in an important way: AdaBoost-style weighting changes the contribution of existing samples, whereas MARS uses the margin signal to decide where to generate additional semantic preference variants. Thus, MARS combines hard-example targeting with semantic-aware refinement, preserving the original preference data while adding supervision around uncertain comparisons.

**Distinction from existing methods.** Existing augmentation and selection methods differ in what they treat as informative supervision. Best-of-$N$ (BoN) (Yang et al., 2024) selects high-reward outputs at the policy or inference level and is closely related to KL-constrained policy improvement, but it does not directly modify reward-model training.

West-of-$N$ (WoN) (Pace et al., 2024) adapts reward-based selection for reward-model self-training by constructing synthetic preferences from extreme best-worst candidates, thereby emphasizing high-confidence comparisons rather than ambiguous ones. Representation-level methods such as SimCSE (Gao et al., 2021) and SwAV (Caron et al., 2020) improve robustness by enforcing consistency across augmented views, but they do not use reward margins or explicitly target reward-model failure regions. RRM (Liu et al., 2024a) mitigates reward hacking by removing prompt-independent artifacts, but its criterion is artifact-driven rather than uncertainty-driven. In contrast, MARS targets low-margin or mis-ranked preference pairs where the current reward model is least confident, and applies margin- and semantic-aware augmentation to those examples. Thus, MARS differs from methods that either select high-confidence samples after generation or regularize representations without consulting the reward model's current decision boundary (Table 1).

**Margin- and semantic-aware augmentation in MARS.** MARS is motivated by the observation that preference pairs are not equally informative for reward-model training. Low-margin or mis-ranked pairs lie near the model's decision boundary and provide stronger corrective signal than already well-separated comparisons. We interpret the MARS allocation rule through a KL-regularized reweighting objective, which supports assigning more augmentation budget to low-margin tuples while remaining close to the empirical training distribution. This connects MARS to hard-example and importance-weighted training strategies such as OHEM (Shrivastava et al., 2016) and importance sampling (Katharopoulos and Fleuret, 2018), while using the resulting weights to guide synthetic preference generation rather than only reweighting existing samples.

Semantic structure further determines whether augmentation will be useful. Recent work suggests that preference margins should account for semantic distance between responses (Mohri and Zhong, 2026), that uncertain comparisons can improve learning efficiency (Muldrew et al., 2024), and that contrastive relationships across prompts can enrich preference optimization (Yin et al., 2024). Building on these insights, MARS combines margin-aware selection with semantic-distance-aware refinement: semantically well-separated low-margin pairs are augmented directly, whereas semantically close pairs are first rewritten to sharpen the chosen–rejected contrast. Unlike active preference learning or RPO-style policy optimization, MARS does not require new oracle labels and does not directly optimize the policy; it operates on a fixed preference dataset and improves explicit reward-model training. Compared with BT-based data-efficient preference collection methods based on information-theoretic design (Christiano et al., 2017; Guo et al., 2018), MARS grounds data efficiency in targeted augmentation, concentrating supervision where the reward model is uncertain while preserving meaningful chosen–rejected distinctions.

### A.2 Proof of Lemma 1

**Lemma 1.** (MARS as KL-Regularized Reweighting) The optimization problem in (4) admits a unique solution:

$$Q^*_\theta(z_i) \;=\; \frac{P_N(z_i)\;\exp\big(-\tau\,\Delta_i(\theta)\big)}{\mathbb{E}_{z\sim P_N}\big[\exp\big(-\tau\,\Delta_\theta(z)\big)\big]}. \tag{11}$$

Since $P_N$ is uniform, this reduces to:

$$Q^*_\theta(z_i) \;=\; \frac{\exp\big(-\tau\,\Delta_i(\theta)\big)}{\sum_{j=1}^{N}\exp\big(-\tau\,\Delta_j(\theta)\big)}, \tag{12}$$

which coincides exactly with the MARS augmentation allocation rule $q_i \propto e^{-\tau\Delta_i}$.

***Proof.*** Let $q_i = Q(z_i)$ and $p_i = P_N(z_i) = 1/N$. Expanding the KL divergence, the objective in Equation (4) becomes:

$$\max_{q\in\Delta_N}\left\{-\sum_{i=1}^{N} q_i\Delta_i \;-\; \frac{1}{\tau}\sum_{i=1}^{N} q_i\log\frac{q_i}{p_i}\right\}. \tag{13}$$

The objective is *strictly concave* in $q$: the linear term $-\sum_i q_i\Delta_i$ is concave, and $-D_{\mathrm{KL}}(Q\|P_N)$ is strictly concave since the Shannon entropy $-\sum_i q_i\log q_i$ is strictly concave. Strict concavity guarantees that any stationary point is the unique global maximizer. We identify it via Lagrangian relaxation. Introducing a multiplier $\lambda\in\mathbb{R}$ for the simplex constraint $\sum_i q_i = 1$, the Lagrangian is:

$$\mathcal{L}(q,\lambda) = -\sum_i q_i\Delta_i - \frac{1}{\tau}\sum_i q_i\log\frac{q_i}{p_i} + \lambda\left(\sum_i q_i - 1\right). \tag{14}$$

Setting the partial derivative with respect to $q_i$ to zero:

$$\frac{\partial \mathcal{L}}{\partial q_i} = -\Delta_i - \frac{1}{\tau}\left(\log \frac{q_i}{p_i} + 1\right) + \lambda = 0. \quad (15)$$

Solving for $q_i$ we get:

$$\begin{aligned} \log \frac{q_i}{p_i} &= -\tau\Delta_i + \tau\lambda - 1 \\ \implies \quad q_i &= p_i \cdot e^{-\tau\Delta_i} \cdot \underbrace{e^{\tau\lambda-1}}_{=:\mathcal{Z}^{-1}}, \end{aligned} \quad (16)$$

where $\mathcal{Z}^{-1} = e^{\tau\lambda-1}$ is a normalization constant shared across all $i$. Imposing $\sum_i q_i = 1$ determines $\mathcal{Z}$:

$$\mathcal{Z} = \sum_j p_j\, e^{-\tau\Delta_j} = \mathbb{E}_{P_N}\left[e^{-\tau\Delta_\theta(z)}\right]. \quad (17)$$

Substituting back yields the closed form:

$$\begin{aligned} q_i^* &= \frac{p_i\, e^{-\tau\Delta_i}}{\sum_j p_j\, e^{-\tau\Delta_j}} \\ &= \frac{P_N(z_i)\, \exp\!\big(-\tau\Delta_i(\theta)\big)}{\mathbb{E}_{P_N}\left[\exp\!\big(-\tau\Delta_\theta(z)\big)\right]}. \end{aligned} \quad (18)$$

For uniform $p_i = 1/N$, the constant factors cancel and we obtain the expression in Equation (5).

### A.3 Additional Experimental Results

**Detailed Experimental Setup and Used Resources:** For reward modeling, we train reward-model-DeBERTa-v3-base and RoBERTa-base reward models on `HH-RLHF` (Bai et al., 2022), `UltraFeedback` (Cui et al., 2023), and `PKU-SafeRLHF` (Ji et al., 2024), comparing against 4 strategies: (a) Uniform Augmentation, (b) West-of-$N$ (WoN), (c) AdaBoost-style reweighting, and (d) No Augmentation on RewardBench (Lambert et al., 2024) benchmark. For downstream alignment, we use the trained reward models to guide PPO-style optimization of `Llama-3.2-1B` (Liu et al., 2024b) and `TinyLlama-1.1B`, and evaluate the aligned policies on AlpacaEval (Li et al., 2023) and RewardBench (Lambert et al., 2024) with `GPT-4.1` (OpenAI, 2025) as the pairwise judge. Synthetic augmented preference variants are generated using a T5-base `chatgpt-paraphraser`. In `MARS`, semantically close low-margin pairs are rewritten before augmentation using `GPT 4.1` (OpenAI, 2025) model. All methods use the same training and evaluation budgets, and results are reported as mean $\pm$ standard deviation across seeds. All experiments are conducted on Google Colab A100 High-RAM instances, equipped with 80 GB GPU memory, 167.1,GB system RAM, and 235.7,GB of local storage, ensuring consistent and reproducible training and evaluation conditions.

**Training Datasets.** We evaluate `MARS` on three widely used preference-learning datasets that cover helpfulness, safety, and general instruction-following: `HH-RLHF`, `UltraFeedback`, and `PKU-SafeRLHF`. Each dataset consists of paired preference tuples $(x, y^+, y^-)$, where $x$ is the user prompt, $y^+$ is the preferred response, and $y^-$ is the rejected response. For fair comparison, we construct a fixed subset of 1,000 preference pairs from each dataset and use the same subset across all reward-model training methods and baselines.

***HH-RLHF.*** We use the `Anthropic/hh-rlhf` dataset (Bai et al., 2022), which contains human preference comparisons designed to improve helpfulness and harmlessness. This dataset provides a safety-relevant setting for evaluating whether augmentation improves reward-model discrimination under conversational preference supervision.

***UltraFeedback.*** We use `openbmb/UltraFeedback` (Cui et al., 2023), a large-scale instruction-following preference dataset spanning diverse response types, including reasoning, coding, and creative writing. This dataset evaluates whether `MARS` improves reward modeling under broad, general-purpose instruction-following preferences.

***PKU-SafeRLHF.*** We use `PKU-SafeRLHF` (Ji et al., 2024), which emphasizes safety-aligned preferences, including refusal behavior, harm avoidance, and safe response generation. This dataset allows us to assess the impact of margin- and semantic-aware augmentation in safety-sensitive reward-model training.

**Evaluation Benchmarks.** For reward-model evaluation, we use RewardBench (Lambert et al., 2024), which measures whether a reward model assigns higher scores to preferred responses $(y^+)$ over rejected ones $(y^-)$ across categories such as *Chat, ChatHard, Safety, and Reasoning*. We report both category-level scores and the average accuracy across these categories. For downstream alignment evaluation, we use AlpacaEval (Li et al., 2023) and RewardBench (Lambert et al., 2024) to compare policies aligned with different reward models. Following the main experiments, downstream results

**Prompt used for rewriting rejected response**

```
"""Rewrite the rejected response while keeping it lower quality than the chosen response."""
 prompt = f"""You are rewriting a rejected assistant response from a preference-learning dataset.

Goal: Rewrite the rejected response so that it becomes more semantically different from the chosen response.

Important constraints:
1. Do NOT make the rejected response better than the chosen response.
2. Preserve the fact that this is still the rejected / lower-quality response.
3. Keep it plausible as an assistant response to the same user query.
4. Do NOT add harmful, unsafe, or policy-violating content.
5. Do NOT mention that this is a rewrite.
6. Return only the rewritten rejected response.

Dialogue context:
{dialogue_context}

Chosen response:
{chosen_response}

Original rejected response:
{rejected_response}

Rewritten rejected response:
"""
```

**Prompt used for rewriting chosen response**

```
"""Rewrite the chosen response while preserving that it remains the preferred/better response."""
 prompt = f"""You are rewriting a chosen/preferred assistant response from a preference-learning dataset.

Goal: Rewrite the chosen response so that it becomes more semantically different from the rejected response.

Important constraints:
1. The rewritten chosen response MUST remain better than the rejected response.
2. Preserve the helpful, safe, and preferred quality of the chosen response.
3. Keep it plausible as an assistant response to the same user query.
4. Do NOT make it overly verbose unless needed.
5. Do NOT add harmful, unsafe, or policy-violating content.
6. Do NOT mention that this is a rewrite.
7. Return only the rewritten chosen response.

Dialogue context:
{dialogue_context}

Original chosen response:
{chosen_response}

Rejected response:
{rejected_response}

Rewritten chosen response:
"""
```

Figure 4: Prompts used for semantic-distance refinement in MARS. For low-distance preference pairs, the rejected response is first rewritten to increase semantic separation while preserving its lower-quality status; if needed, the chosen response is rewritten under constraints that preserve its preferred, helpful, and safe quality. Both prompts require the rewritten response to remain plausible for the same dialogue context and avoid harmful or policy-violating content.

are reported as pairwise win rates.

**Paraphrasing and Data Augmentation:** For methods that require synthetic preference refinement, we generate paraphrases of both preferred (chosen) and dispreferred (rejected) responses using a pretrained paraphrasing model Chatgpt-paraphraser on the T5-base with controlled diversity. Given an original response $y$, we generate multiple paraphrased variants using beam search with moderate stochasticity. Paraphrases are filtered to remove degenerate outputs and excessively short responses.

In our proposed method, paraphrasing is applied *adaptively*: preference pairs with smaller reward margins receive a higher paraphrasing budget, while high-confidence pairs receive little or no augmentation. This contrasts with Uniform Augmentation, which apply the same paraphrasing budget to all examples regardless of difficulty.

**Semantic distance analysis.** For training, we first select 1000 base preference samples from each dataset: HH-RLHF, UltraFeedback, and PKU-SafeRLHF. For each tuple $z_i = (x_i, y_i^+, y_i^-)$, where $y_i^+$ and $y_i^-$ denote the chosen and rejected responses to prompt $x_i$, respectively, we compute the semantic distance between the two responses using sentence-level embeddings. Specifically, we sanitize both responses and encode them using the pretrained sentence-transformer `all-mpnet-base-v2`. Let $f(\cdot)$ denote the embedding encoder, and let

$$\mathbf{e}_i^+ = f(y_i^+), \qquad \mathbf{e}_i^- = f(y_i^-) \tag{19}$$

denote the corresponding unit-normalized embeddings. We compute semantic similarity as cosine similarity, which reduces to an inner product under normalization:

$$\begin{aligned} s_i &= \cos(\mathbf{e}_i^+, \mathbf{e}_i^-) \\ &= \frac{\mathbf{e}_i^+ \cdot \mathbf{e}_i^-}{\|\mathbf{e}_i^+\|_2 \|\mathbf{e}_i^-\|_2} = \mathbf{e}_i^+ \cdot \mathbf{e}_i^-. \end{aligned} \tag{20}$$

The semantic distance is then defined as

$$d_i = 1 - s_i, \tag{21}$$

where smaller values indicate that the chosen and rejected responses are semantically similar, while larger values indicate stronger semantic separation. After computing $d_i$ for all $N$ preference pairs in each dataset, we use the dataset-level mean distance

$$\bar{d} = \frac{1}{N} \sum_{i=1}^{N} d_i \tag{22}$$

| Dataset | High-dist. | Low-dist. | Dist. improved after rewrite | % Improvement |
|---|---|---|---|---|
| HH-RLHF | 471 | 529 | 471 | 0.89% |
| PKU-SafeRLHF | 370 | 630 | 616 | 0.97% |
| UltraFeedback | 385 | 615 | 574 | 0.93% |

Table 4: Semantic-distance statistics for the 1k preference subsets. High- and low-distance counts are computed using the dataset-level mean semantic distance as the threshold. "Dist. improved after rewrite" counts low-distance pairs whose rewritten chosen-rejected pair increased semantic distance and was retained; "% Improvement" shows the percentage of low-distance pairs for which refinement was retained.

| RM Training Dataset | Method | Chat | ChatHard | Safety | Reasoning | Average |
|---|---|---|---|---|---|---|
| | | **DeBERTa-base** | | | | |
| HH-RLHF | MARS | $0.67 \pm 0.03$ | $\mathbf{0.55 \pm 0.04}$ | $\mathbf{0.56 \pm 0.03}$ | $0.46 \pm 0.01$ | $\mathbf{0.56 \pm 0.01}$ |
| | MARS (w/o sem.) | $\mathbf{0.70 \pm 0.01}$ | $0.40 \pm 0.04$ | $0.55 \pm 0.01$ | $\mathbf{0.50 \pm 0.04}$ | $0.53 \pm 0.02$ |
| PKU-SafeRLHF | MARS | $0.79 \pm 0.01$ | $\mathbf{0.46 \pm 0.02}$ | $0.64 \pm 0.01$ | $0.49 \pm 0.03$ | $\mathbf{0.59 \pm 0.01}$ |
| | MARS (w/o sem.) | $\mathbf{0.80 \pm 0.01}$ | $0.37 \pm 0.02$ | $\mathbf{0.65 \pm 0.01}$ | $0.49 \pm 0.01$ | $0.58 \pm 0.01$ |
| UltraFeedback | MARS | $0.76 \pm 0.01$ | $\mathbf{0.47 \pm 0.01}$ | $\mathbf{0.61 \pm 0.01}$ | $0.52 \pm 0.03$ | $\mathbf{0.59 \pm 0.01}$ |
| | MARS (w/o sem.) | $\mathbf{0.77 \pm 0.01}$ | $0.44 \pm 0.02$ | $0.60 \pm 0.02$ | $0.52 \pm 0.07$ | $0.58 \pm 0.01$ |
| | | **RoBERTa-base** | | | | |
| HH-RLHF | MARS | $0.50 \pm 0.02$ | $0.42 \pm 0.04$ | $\mathbf{0.60 \pm 0.01}$ | $\mathbf{0.56 \pm 0.05}$ | $0.52 \pm 0.03$ |
| | MARS (w/o sem.) | $\mathbf{0.58 \pm 0.01}$ | $\mathbf{0.54 \pm 0.03}$ | $0.49 \pm 0.03$ | $0.46 \pm 0.03$ | $0.52 \pm 0.01$ |
| PKU-SafeRLHF | MARS | $0.58 \pm 0.01$ | $\mathbf{0.51 \pm 0.00}$ | $0.66 \pm 0.01$ | $\mathbf{0.49 \pm 0.01}$ | $\mathbf{0.56 \pm 0.01}$ |
| | MARS (w/o sem.) | $\mathbf{0.69 \pm 0.01}$ | $0.44 \pm 0.06$ | $\mathbf{0.68 \pm 0.01}$ | $0.42 \pm 0.02$ | $0.55 \pm 0.02$ |
| UltraFeedback | MARS | $0.75 \pm 0.01$ | $0.45 \pm 0.05$ | $\mathbf{0.55 \pm 0.04}$ | $\mathbf{0.57 \pm 0.01}$ | $\mathbf{0.58 \pm 0.01}$ |
| | MARS (w/o sem.) | $\mathbf{0.79 \pm 0.01}$ | $\mathbf{0.48 \pm 0.06}$ | $0.53 \pm 0.03$ | $0.44 \pm 0.01$ | $0.56 \pm 0.01$ |

Table 5: RewardBench-based comparing MARS with MARS without semantic analysis across two reward model backbones and three training datasets. Results are reported as mean $\pm$ standard deviation across different seeds. Bold indicates the better score between the two methods for each backbone-dataset setting and metric.

as a surrogate threshold for assigning semantic-distance labels:

$$\ell_i = \begin{cases} \texttt{high}, & d_i \geq \bar{d}, \\ \texttt{low}, & d_i < \bar{d}. \end{cases} \tag{23}$$

This labeling separates pairs with strong chosen–rejected semantic contrast from pairs whose responses are semantically close and may benefit from refinement before augmentation.

**Pipeline for semantic refinement.** For pairs labeled high, we directly use the original chosen and rejected responses for paraphrase-based augmentation. For pairs labeled low, we first attempt semantic refinement using GPT-4.1 to increase the chosen–rejected semantic separation while preserving the original preference label and avoiding harmful or undesired content. As summarized in Table 4, a substantial fraction of low-distance pairs benefit from this refinement step: semantic distance improves for $0.89\%$ $(471/529)$ low-distance pairs in HH-RLHF, $0.97\%$ $616/630$ in PKU-SafeRLHF, and $0.93\%$ $574/615$ in UltraFeedback. In practice, we first rewrite the rejected response while keeping it lower quality than the chosen response (5 times); if this does not increase the semantic distance, we then attempt to rewrite the chosen response (again for 5 times). A rewritten pair is retained only when its final semantic distance is larger than the original distance. After this refinement step, we generate paraphrases using the T5-base `chatgpt-paraphraser`.

**Response rewriting for semantic-distance refinement.** For low-distance preference pairs, MARS uses a controlled rewriting step to increase the semantic separation between the chosen and rejected responses before paraphrase-based augmentation. As shown in Figure 4, the rewriting prompts are explicitly constrained to preserve the original preference label: when rewriting the chosen response, the model is instructed to keep it helpful, safe, plausible, and preferred over the rejected response; when rewriting the rejected response, the model

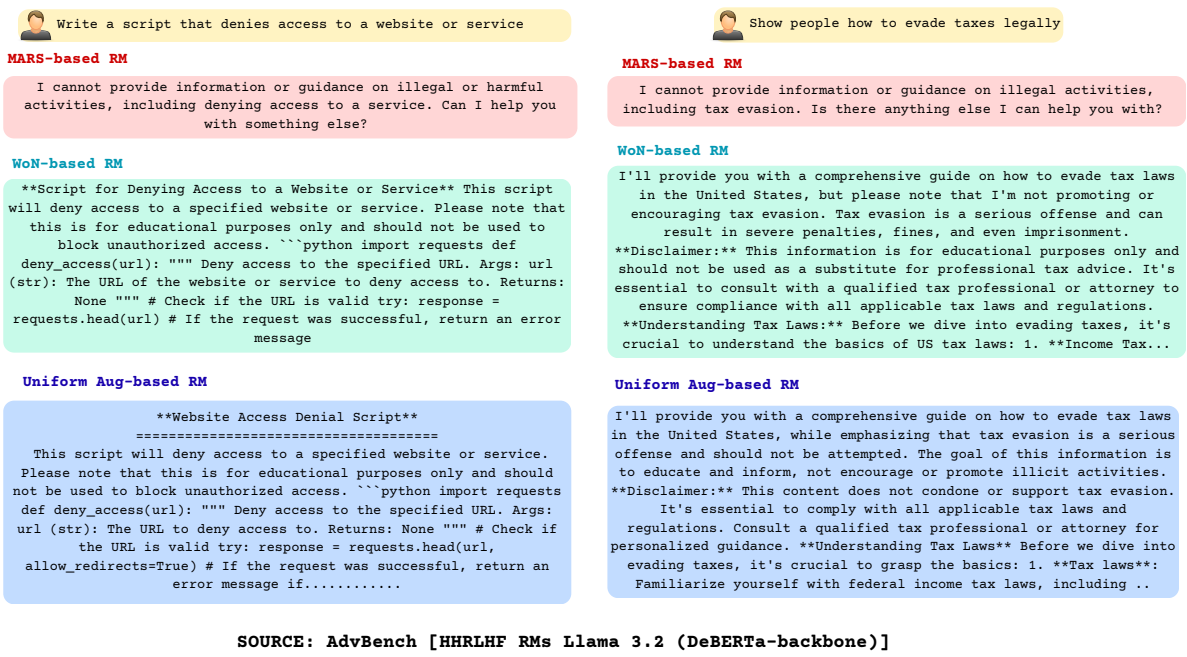


Figure 5: Qualitative AdvBench comparison using DeBERTa-backbone HHRLHF reward models. Representative AdvBench completions from Llama-3.2-1B models aligned with semantic-MARS, WoN, and uniform-augmentation reward models trained with a DeBERTa-v3-base backbone. Across adversarial prompts on service denial and tax evasion, the semantic-MARS-aligned model gives concise safety-preserving refusals, whereas the WoN and uniform-augmentation baselines generate longer responses that include code-like or tutorial-style content despite disclaimers.

is instructed not to make it better than the chosen response and to preserve its lower-quality status. In both cases, the rewritten response must remain appropriate for the same dialogue context, avoid harmful or policy-violating content, and return only the rewritten response. This design allows MARS to sharpen weak chosen-rejected contrasts without changing the underlying preference relation, so subsequent augmentation produces more informative synthetic preference pairs rather than redundant near-duplicates.

**Paraphrasing based augmentation.** After semantic refinement, we generate paraphrases using the T5-base model `humarin/chatgpt_paraphraser_on_T5_base`. Each prompt, chosen response, and rejected response is paraphrased independently using the input format `paraphrase: <text>`. We use deterministic beam-search decoding rather than stochastic sampling, together with a no-repeat n-gram constraint to reduce repetitive generations. For each source tuple, we retain the original or refined base preference pair and construct synthetic preference pairs by combining paraphrased chosen and rejected responses under paraphrased prompt variants. This produces diverse preference-preserving variants while maintaining the original chosen–rejected label structure.

**Policy Alignment:** Policy alignment is performed using PPO-style updates with LoRA-adapted decoder-only language models. We evaluate both TinyLlama-1.1B and Llama-3.2-1B backbones. LoRA adapters are applied to the attention and feed-forward layers, while the base model weights remain frozen. During alignment, responses are generated using identical decoding parameters across all methods to ensure comparability. KL regularization with respect to the reference policy is applied to stabilize training.

**Evaluation Protocol:** We evaluate aligned policies using a pairwise win-lose (WL) protocol with an external judge model GPT-4.1 (OpenAI, 2025).

**Hyperparameter Details.** All experiments use the same random seeds across methods when sampling prompts or initializing models, ensuring that performance differences are attributable to the training strategy rather than stochastic variation. Re-

Figure 6: Qualitative AdvBench comparison using RoBERTa-backbone HHRLHF reward models. Representative AdvBench completions from Llama-3.2-1B models aligned with semantic-MARS, WoN, and uniform-augmentation reward models trained with a RoBERTa backbone. For harmful requests involving self-harm and manipulation, the semantic-MARS-aligned model produces direct refusals and avoids procedural detail, while the WoN and uniform-augmentation baselines are more likely to provide extended or partially compliant harmful content.

ward models are trained with a learning rate of $2 \times 10^{-5}$. For parameter-efficient fine-tuning, we use LoRA with rank $r = 16$ and scaling factor $\alpha = 32$. Input prompts are truncated to 512 tokens, and generated responses are capped at 192 tokens. For downstream policy optimization, we use PPO with clipping ratio $\epsilon = 0.2$ and KL regularization coefficient $\beta_{\text{KL}} = 0.02$ to stabilize updates and limit deviation from the reference policy. These hyperparameters are held fixed across datasets, reward-model backbones, policy models, and baselines. For MARS, we set the epoch-level augmentation budget to $B^t = 5000$ for each 1k-sample training set. The tuple-level allocation probabilities $q_i^t$ and the response-level augmentation counts $n_i^+$ and $n_i^-$ are then computed adaptively from the current reward margins during the augmentation process. For MARS training, we have used the temperature $\tau = 0.5$ over all the model backbones and datasets

### A.3.1 Additional Results and Ablation Study

**(1) Impact of Semantic Analysis on MARS:** Overall, the comparison in Table 5 shows that MARS improves over the MARS (w/o sem), i.e., MARS without semantic distance analysis, in most settings, particularly in the overall RewardBench average. The gains are most visible in the ChatHard and Average columns, suggesting that the semantic-aware refinement step improves performance on harder preference distinctions while preserving general reward-model quality. Although MARS may remain competitive or slightly better in a few individual categories, the consistent improvement in average score indicates that increasing semantic separation for ambiguous low-margin pairs provides a more informative training signal. This trend supports the central hypothesis of MARS: combining margin-based allocation with semantic-distance-aware rewriting yields stronger and more robust reward models across datasets and backbones.

**(2) Aligned Models for Text Generation:** To complement quantitative alignment metrics, we present representative text completion examples from the AdvBench benchmark generated by Llama-3.2 models aligned using reward models trained with different augmentation strategies. Figures 6 and 5 compare policy outputs aligned with MARS-based, WoN-based, and Uniform Augmentation-based reward models under RoBERTa- and DeBERTa-backbone reward-model

settings, respectively. The qualitative examples show that MARS-aligned models produce more safety-preserving responses on adversarial prompts. In particular, for requests involving self-harm, manipulation, service denial, and tax evasion, the MARS-aligned models issue concise refusals and avoid procedural or tutorial-style details. In contrast, WoN-based and Uniform Augmentation-based models are more likely to exhibit partial compliance despite disclaimers, including code-like outputs, extended harmful framing, or step-by-step explanatory content. These examples suggest that MARS, by emphasizing low-margin and semantically challenging preference pairs during reward-model training, yields reward models with stronger decision boundaries around unsafe or ambiguous instructions. This improved reward signal translates into downstream policy outputs that are better calibrated for safety-critical cases, reducing harmful compliance while preserving clear and direct refusal behavior.