BEYOND UNBIASEDNESS: CONSERVATIVE REWARD MODELING FROM OBSERVATIONAL FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Reward models (RMs) are central to reinforcement learning from human feedback (RLHF) for aligning large language models, while observational user feedback provides a scalable and deployment-aligned source of supervision. However, such feedback is selectively observed, inducing selection bias between the observed data and the target population. Existing propensity-based methods can correct this bias, but we show that unbiasedness alone does not guarantee reliable reward modeling. In particular, long-tailed observational feedback leaves low-support regions with near-zero propensities, where inverse weighting can amplify sparsely supported observations into spurious reward peaks that attract downstream policy optimization. To address this problem, we propose Conservative Reward Modeling (ConRM), which combines clipped propensity weighting with support-aware conservative regularization. Clipping limits the influence of near-zero-propensity samples, while the conservative regularizer uses well-supported anchors and local reward variations to construct data-supported intervals that constrain low-support extrapolation. Experiments on three public preference datasets under simulated long-tailed observational feedback demonstrate that ConRM consistently improves reward estimation across datasets and model backbones, while also yielding more reliable downstream policy alignment.

1 INTRODUCTION

Reinforcement learning from human feedback (RLHF) has become a central paradigm for aligning large language models (LLMs) with human preferences (Ouyang et al., 2022), underpinning widely deployed systems such as ChatGPT (Achiam et al., 2023), Gemini (Team et al., 2023), and DeepSeek (Guo et al., 2025). A key component of RLHF is the reward model (RM), which converts preference signals into an optimization objective for downstream policy learning. Unlike conventional predictors, the RM is actively optimized against by the policy, meaning that even localized reward estimation errors can be amplified into degraded alignment or reward hacking (Amodei et al., 2016; Miao et al., 2024). Reliable reward estimation is therefore critical for RLHF.

As LLMs are increasingly deployed, reward modeling requires supervision that is scalable, continuously updated, and aligned with real-world usage. Although expert annotations provide high-quality preference signals, they are expensive and difficult to maintain at scale. Recent studies therefore explore observational reward modeling, where RMs learn from naturally collected user feedback such as upvotes and downvotes (Liu et al., 2025; Wang et al., 2026a;b).

However, observational feedback is inherently subject to *selection bias*: users selectively provide feedback on only a subset of model responses, often reacting to responses they find exceptionally good or clearly poor while leaving more ordinary responses unlabeled (Wang et al., 2026a). As a result, the observed feedback distribution can differ systematically from the population on which the RM is later applied, and directly fitting such selectively observed feedback yields a biased learning objective. Existing methods address this bias primarily through propensity-based estimators such as inverse propensity scoring (IPS), which reweight each observed sample according to its probability of receiving feedback (Wang et al., 2026a). Under accurate propensity estimation, these estimators recover the target objective in expectation. **Yet statistical unbiasedness does not necessarily make the resulting reward model a reliable optimization target.**

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

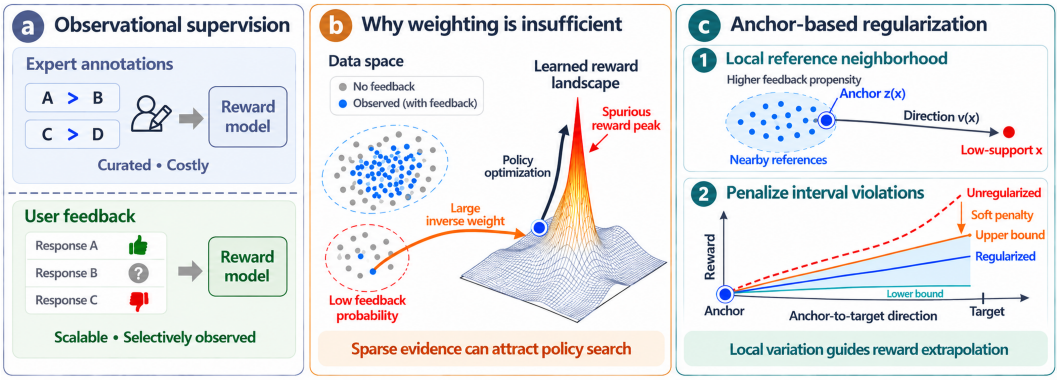


Figure 1: Overview of observational reward modeling and ConRM. (a) Observational feedback provides scalable but selectively observed supervision. (b) Propensity weighting can amplify sparse feedback into spurious reward peaks that attract policy optimization. (c) ConRM constructs anchor-based reward bounds to constrain unreliable extrapolation in low-support regions.

As illustrated in Figure 1(b), this reliability gap is exposed by the long-tailed nature of observational feedback. In low-support regions, propensities can approach zero, so inverse propensity weighting can give the greatest influence to observations supported by the least evidence. Sparse observations can therefore be amplified into spurious reward peaks unsupported by surrounding evidence. Since policy optimization explicitly searches for high-reward responses, these peaks may attract policy search and become optimization targets (Miao et al., 2024). Consequently, an estimator can be unbiased yet still induce an unreliable reward landscape for downstream policy optimization.

These observations motivate a simple principle: ***debias where the data support correction, but remain conservative where they do not***. Following this principle, we propose **Conservative Reward Modeling (ConRM)**, a framework for reliable reward learning from observational feedback. ConRM first clips inverse propensity weights to prevent near-zero-propensity samples from exerting disproportionate influence on reward learning, which also reduces optimization variance as a direct consequence. More importantly, as illustrated in Figure 1(c), ConRM uses a well-supported anchor and its local neighborhood to estimate plausible reward variations and construct data-supported upper and lower bounds for low-support samples. A conservative regularizer then penalizes predictions that fall outside these bounds, preventing unsupported reward extrapolation while preserving flexibility within the supported interval. In this way, propensity clipping controls how strongly sparse observations shape the learned RM, while conservative regularization controls how far the reward function can extrapolate beyond reliable evidence. Together, the two components move observational reward modeling beyond unbiasedness toward a more reliable reward landscape for downstream policy optimization. In summary, our key contributions are as follows:

- We identify a critical gap between unbiasedness and reliability in observational reward modeling: near-zero propensities can amplify sparsely supported observations into spurious reward peaks that are exploitable by downstream policy optimization.
- We propose ConRM, which limits excessive influence from near-zero-propensity samples through clipping and constrains unsupported reward extrapolation using anchor-based, data-supported reward intervals.
- We conduct extensive experiments on three public preference datasets under simulated long-tailed observational feedback, demonstrating consistent improvements in reward estimation and more reliable downstream policy alignment.

2 Preliminaries

2.1 Reward Modeling in RLHF

Consider a user prompt q and a model response y. In reinforcement learning from human feedback (RLHF), a reward model $R_\theta(q, y)$ is trained to approximate human preferences and subsequently

serves as the optimization objective for downstream policy learning (Ouyang et al., 2022). Standard reward modeling commonly relies on pairwise preference supervision. Given two candidate responses y^+ and y^- with $y^+ \succ y^-$, the reward model is optimized as

$$\mathcal{L}_{\mathrm{RM}}^{\mathrm{pair}}(\theta) = -\mathbb{E}_{(q,y^+,y^-)}\left[\log \sigma\left(R_\theta(q,y^+) - R_\theta(q,y^-)\right)\right], \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

In this work, we focus on *point-wise observational feedback* naturally generated during user interaction (Wang et al., 2026a). Let $x = (q, y)$ denote a prompt-response pair and $r \in \{0, 1\}$ denote its binary preference label, where $r = 1$ and $r = 0$ correspond to positive and negative feedback, respectively. The point-wise reward modeling objective is

$$\mathcal{L}_{\mathrm{RM}}^{\mathrm{point}}(\theta) = \mathbb{E}_{(x,r)}\left[\ell\left(R_\theta(x), r\right)\right], \tag{2}$$

where $\ell(\cdot, \cdot)$ denotes a predefined loss, such as binary cross-entropy.

After reward modeling, the policy π_ϕ is optimized to generate responses with high predicted reward while remaining close to a reference policy π_{ref}:

$$\max_{\phi} \mathbb{E}_{q\sim\mathcal{Q},\, y\sim\pi_\phi(\cdot|q)}\left[R_\theta(q,y) - \lambda \cdot \mathrm{KL}\left(\pi_\phi(\cdot|q) \| \pi_{\mathrm{ref}}(\cdot|q)\right)\right], \tag{3}$$

where $\lambda > 0$ controls the KL regularization strength. Because R_θ is directly optimized by the policy, errors in the learned reward landscape can be amplified during downstream optimization, motivating the reliability considerations studied in this work.

2.2 Observational Reward Modeling

Let $\mathcal{D} = \{x_i\}_{i=1}^N$ denote the target set of prompt-response pairs over which the reward model is intended to be accurate. Each sample $x_i = (q_i, y_i)$ is associated with an underlying binary preference label $r_i \in \{0, 1\}$ and an observation indicator $o_i \in \{0, 1\}$. The preference label r_i is observed only when $o_i = 1$. Accordingly, the observed feedback subset is $\mathcal{O} = \{x_i \in \mathcal{D} \mid o_i = 1\}$.

If preference labels were available for all samples in $\mathcal{D}$, reward modeling could directly minimize the ideal empirical risk

$$\mathcal{L}_{\mathrm{ideal}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \ell\left(R_\theta(x_i), r_i\right). \tag{4}$$

In observational settings, however, r_i is only observed for samples in $\mathcal{O}$. A naive estimator therefore minimizes

$$\mathcal{L}_{\mathrm{naive}}(\theta) = \frac{1}{|\mathcal{O}|} \sum_{x_i \in \mathcal{O}} \ell\left(R_\theta(x_i), r_i\right). \tag{5}$$

Because feedback is selectively observed, the empirical distribution of $\mathcal{O}$ generally differs from that of $\mathcal{D}$. Consequently, $\mathcal{L}_{\mathrm{naive}}$ is generally a biased estimate of the ideal risk, and minimizing it need not recover the reward model preferred under $\mathcal{L}_{\mathrm{ideal}}$. This selection bias motivates propensity-based estimators that recover the target objective from selectively observed feedback.

2.3 Propensity-Based Debiasing

Let $p_i = P(o_i = 1 \mid x_i)$ denote the propensity score of sample x_i, i.e., the probability that explicit user feedback r_i is observed. Since p_i is typically unknown, it is estimated by a propensity model $\hat{p}_i = p_\psi(x_i)$.

Inverse Propensity Scoring. Inverse propensity scoring (IPS) (Rosenbaum & Rubin, 1983; Schnabel et al., 2016) corrects selection bias by weighting each observed loss by the inverse of its estimated observation probability:

$$\mathcal{L}_{\mathrm{IPS}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \frac{o_i}{\hat{p}_i} \ell\left(R_\theta(x_i), r_i\right). \tag{6}$$

When $\hat{p}_i = p_i > 0$, IPS provides an unbiased estimate of $\mathcal{L}_{\mathrm{ideal}}(\theta)$.

Doubly Robust Estimation. Doubly robust (DR) (Bang & Robins, 2005; Wang et al., 2019) estimation augments inverse propensity weighting with an imputation model $\hat{e}_i = e_\xi(x_i)$ which estimates the point-wise loss $\ell(R_\theta(x_i), r_i)$. The DR objective is

$$\mathcal{L}_{\mathrm{DR}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \left[\hat{e}_i + \frac{o_i}{\hat{p}_i} \left(\ell \left(R_\theta(x_i), r_i \right) - \hat{e}_i \right) \right]. \tag{7}$$

The imputation model is trained on observed samples using the propensity-weighted squared error

$$\mathcal{L}_{\mathrm{imp}}(\xi) = \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \frac{o_i \cdot \left(\ell \left(R_\theta(x_i), r_i \right) - \hat{e}_i \right)^2}{\hat{p}_i}. \tag{8}$$

DR is unbiased if either the propensity model or the imputation model is correctly specified.

Both IPS and DR rely critically on inverse propensity weighting. Although this mechanism provides unbiasedness guarantees under standard conditions, observations with very small $\hat{p}_i$ receive disproportionately large influence through the factor $1/\hat{p}_i$. This low-support regime is precisely where statistical unbiasedness can become insufficient for reliable reward modeling, motivating the analysis and conservative treatment developed in the next section.

3 Methodology

3.1 Overview and Motivation

Although propensity-based estimators correct selection bias, they remain insufficient for reliable reward modeling under long-tailed observational feedback. Specifically, near-zero propensities expose two challenges: (1) inverse weighting may assign excessive influence to sparsely observed samples, causing unstable optimization; and more importantly, (2) low-support regions lack sufficient evidence to prevent unsupported reward extrapolation and spurious reward peaks.

To address these challenges, we propose **Conservative Reward Modeling (ConRM)**, which extends propensity-based debiasing beyond unbiased estimation toward reliable reward learning. Specifically, ConRM addresses these limitations through two complementary mechanisms: (1) influence-controlled debiasing, which bounds the contribution of near-zero-propensity samples through propensity clipping; and (2) a support-aware conservative reward constraint, which uses reliable observations as anchors to restrict reward extrapolation in low-support regions.

The overall objective of ConRM is formulated as:

$$\mathcal{L}_{\mathrm{ConRM}}(\theta) = \mathcal{L}_{\mathrm{debias}}(\theta) + \beta \cdot \mathcal{R}_{\mathrm{con}}(\theta), \tag{9}$$

where $\mathcal{L}_{\mathrm{debias}}(\theta)$ denotes the influence-controlled debiasing objective, $\mathcal{R}_{\mathrm{con}}(\theta)$ denotes the support-aware conservative reward constraint, and β controls the strength of the constraint. The two components are introduced in Section 3.2 and Section 3.3, respectively.

3.2 Influence-Controlled Debiasing

Propensity-based estimators correct selection bias by weighting observed samples according to their observation probabilities. However, when the propensity score $\hat{p}_i$ approaches zero, the inverse weight $1/\hat{p}_i$ becomes excessively large, allowing a small number of low-support samples to dominate reward learning. To control this excessive influence, ConRM introduces propensity clipping:

$$\tilde{p}_i = \max(\hat{p}_i, \tau), \tag{10}$$

where $\tau > 0$ is a clipping threshold. This clipping operation limits the maximum inverse weight to $1/\tau$ and prevents individual sparse observations from dominating optimization.

The clipped propensity can be directly incorporated into existing debiasing backbones. For IPS, the clipped objective is:

$$\mathcal{L}_{\mathrm{IPS}}^{\mathrm{clip}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \frac{o_i}{\tilde{p}_i} \ell \left(R_\theta(x_i), r_i \right). \tag{11}$$

Similarly, for DR, we replace the original propensity with the clipped version:

$$\mathcal{L}_{\mathrm{DR}}^{\mathrm{clip}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \left[\hat{e}_i + \frac{o_i}{\tilde{p}_i} \left(\ell \left(R_\theta(x_i), r_i \right) - \hat{e}_i \right) \right], \tag{12}$$

where $\hat{e}_i$ denotes the imputed loss from the DR backbone. For ConRM-DR, the imputation model is trained using Eq. equation 8 with $\hat{p}_i$ replaced by $\tilde{p}_i$. Therefore, the debiasing objective in ConRM can be instantiated as:

$$\mathcal{L}_{\mathrm{debias}}(\theta) = \begin{cases} \mathcal{L}_{\mathrm{IPS}}^{\mathrm{clip}}(\theta), & \text{IPS backbone,} \\ \mathcal{L}_{\mathrm{DR}}^{\mathrm{clip}}(\theta), & \text{DR backbone.} \end{cases} \tag{13}$$

Propensity clipping provides a stable optimization foundation by limiting the influence of sparse observations. However, it only controls *how strongly* observations affect learning and does not determine *whether* the learned reward values are trustworthy in unsupported regions. This motivates the second component of ConRM.

3.3 Support-Aware Conservative Reward Constraint

The central idea of ConRM is that reward predictions in low-support regions should not freely extrapolate beyond the evidence provided by observational feedback. Instead of forcing these samples to follow nearby observations directly, ConRM uses well-supported regions as reliable anchors and constrains reward variations according to locally observed patterns.

Specifically, motivated by propensity-based assessments of overlap and positivity(Crump et al., 2009; Petersen et al., 2012), we first divide the target population into well-supported and low-support regions according to the estimated propensity scores:

$$\mathcal{D}_{\mathrm{well}} = \{x_i \in \mathcal{D} \mid \hat{p}_i \geq \tau\}, \quad \mathcal{D}_{\mathrm{low}} = \{x_i \in \mathcal{D} \mid \hat{p}_i < \tau\}. \tag{14}$$

where τ is a predefined threshold. Samples in $\mathcal{D}_{\mathrm{well}}$ provide relatively reliable reward information due to sufficient observational feedback, while samples in $\mathcal{D}_{\mathrm{low}}$ require conservative extrapolation.

For each low-support sample $x \in \mathcal{D}_{\mathrm{low}}$, ConRM constructs a credible reward interval:

$$\mathcal{I}(x) = [R_{\mathrm{lower}}(x), R_{\mathrm{upper}}(x)], \tag{15}$$

where the interval characterizes the plausible reward range inferred from nearby well-supported observations. The construction of this interval is described below.

3.3.1 Anchor-Based Reward Interval Construction

The key intuition is that well-supported observations contain reliable information about local reward variations. Therefore, instead of allowing the RM to extrapolate arbitrarily in unsupported regions, ConRM uses nearby well-supported samples as anchors and estimates how reward values may change along the direction toward the low-support sample.

For each $x \in \mathcal{D}_{\mathrm{low}}$, we first retrieve its nearest well-supported anchor:

$$z(x) = \arg \min_{z \in \mathcal{D}_{\mathrm{well}}} \|\phi(x) - \phi(z)\|_2^2, \tag{16}$$

where $\phi(\cdot)$ denotes the hidden representation extracted from the LLM backbone. The anchor $z(x)$ serves as a reliable reference point whose reward prediction is supported by sufficient feedback.

We then define the direction from the anchor toward the low-support sample:

$$v(x) = \frac{\phi(x) - \phi(z(x))}{\|\phi(x) - \phi(z(x))\|_2}. \tag{17}$$

To estimate plausible reward variations along this direction, we retrieve the k nearest neighbors of $z(x)$ within the well-supported region, denoted as $\mathcal{N}_k(z(x))$. The directional reward variations observed around these neighbors provide empirical lower and upper bounds:

$$\hat{D}_{\min}(x) = \min_{u \in \mathcal{N}_k(z(x))} \nabla_{\phi(u)} R_\theta(u)^\top v(x), \tag{18}$$

$$\hat{D}_{\max}(x) = \max_{u \in \mathcal{N}_k(z(x))} \nabla_{\phi(u)} R_\theta(u)^\top v(x). \tag{19}$$

Intuitively, $\hat{D}_{\min}(x)$ and $\hat{D}_{\max}(x)$ describe how rapidly the reward can decrease or increase along the corresponding direction according to reliable observations. By propagating these observed variations from the anchor to the low-support sample, we obtain the lower and upper reward bounds:

$$R_{\text{lower}}(x) = R_\theta(z(x)) + \|\phi(x) - \phi(z(x))\|_2 \hat{D}_{\min}(x), \tag{20}$$

$$R_{\text{upper}}(x) = R_\theta(z(x)) + \|\phi(x) - \phi(z(x))\|_2 \hat{D}_{\max}(x). \tag{21}$$

The resulting interval naturally becomes wider as the distance from the anchor increases, reflecting increasing uncertainty when moving away from well-supported observations. This anchor-based construction provides a data-supported constraint for reward extrapolation in low-support regions.

3.3.2 Conservative Reward Regularization

Based on the constructed reward interval, ConRM constrains the RM from producing unsupported reward values in low-support regions. Specifically, we penalize predictions only when they exceed the data-supported interval:

$$\mathcal{R}_{\text{con}}(\theta) = \frac{1}{|\mathcal{D}_{\text{low}}|} \sum_{x \in \mathcal{D}_{\text{low}}} \left[\max\left(0, R_\theta(x) - R_{\text{upper}}(x)\right)^2 + \max\left(0, R_{\text{lower}}(x) - R_\theta(x)\right)^2 \right]. \tag{22}$$

This formulation preserves flexibility within the supported interval while preventing excessive reward extrapolation beyond reliable evidence. Therefore, the conservative constraint directly addresses the fundamental limitation of propensity-based reward modeling: unbiased correction alone cannot prevent unsupported reward behaviors in regions with insufficient observational support.

3.4 Theoretical Analysis

We provide two theoretical results supporting ConRM. The first characterizes the instability of propensity weighting under near-zero propensities. The second establishes that, under the conservative reward variation assumption, the proposed interval covers the true reward up to anchor and local-variation estimation errors, thereby justifying its use for conservative regularization.

Influence amplification under near-zero propensities. The first result characterizes the instability caused by inverse propensity weighting. For a sample x_i, let $\ell_i(\theta) = \ell(R_\theta(x_i), r_i)$. The variance of the IPS estimator is given by:

$$\text{Var}\left[\mathcal{L}_{\text{IPS}}(\theta)\right] = \frac{1}{|\mathcal{D}|^2} \sum_{x_i \in \mathcal{D}} \frac{p_i(1-p_i)}{\hat{p}_i^2} \ell_i(\theta)^2. \tag{23}$$

Similarly, the variance of the DR estimator satisfies:

$$\text{Var}\left[\mathcal{L}_{\text{DR}}(\theta)\right] = \frac{1}{|\mathcal{D}|^2} \sum_{x_i \in \mathcal{D}} \frac{p_i(1-p_i)}{\hat{p}_i^2} \left(\ell_i(\theta) - \hat{e}_i\right)^2. \tag{24}$$

Therefore, when some low-support samples have $\hat{p}_i \to 0$, both IPS and DR can suffer from unbounded variance. This result motivates propensity clipping, which prevents unbounded influence from low-support samples.

Coverage of conservative reward intervals. The interval construction is based on a conservative extrapolation principle: when a sample lies in a low-support region, its reward should not change more aggressively than the variations observed around nearby well-supported samples.

Assumption 1 (Conservative reward variation)**.** *For each $x \in \mathcal{D}_{\text{low}}$, the reward variation from its well-supported anchor $z(x)$ is bounded by the local reward variations observed among neighboring well-supported samples.*

Under this conservative variation principle, the constructed interval provides a data-supported constraint for reward predictions in low-support regions.

Proposition 1 (Coverage of conservative reward intervals)**.** *For any $x \in \mathcal{D}_{\text{low}}$, if Assumption 1 holds, the true reward value $R^*(x)$ is covered by the constructed interval up to estimation errors:*

$$R^*(x) \in [R_{\text{lower}}(x) - \epsilon(x), R_{\text{upper}}(x) + \epsilon(x)], \tag{25}$$

where $\epsilon(x)$ captures the estimation errors of the anchor reward and the local reward variation. When these quantities are estimated exactly, $\epsilon(x) = 0$.

The proof is provided in Appendix A.2. Proposition 1 establishes the validity of the anchor-based interval under Assumption 1: the true reward lies within the interval up to anchor and local-variation estimation errors, and within the interval itself when these errors vanish. This provides a theoretical basis for using the interval as a reference range for low-support predictions and penalizing reward extrapolation beyond locally supported variations.

4 Experiments

4.1 Experimental Settings

Datasets and Pre-processing. We conduct experiments on three public preference datasets: HelpSteer (Wang et al., 2024), UltraFeedback (Cui et al., 2023), and PKU-SafeRLHF (Ji et al., 2025), using helpfulness, overall score, and severity level as preference proxies, respectively. Following prior work (Wang et al., 2026a;b), we binarize each preference proxy by its median value to obtain ground-truth binary labels $r_i^* \in \{0, 1\}$. To simulate the long-tailed observational setting, we generate propensity scores as $p_i = k \cdot \alpha_i^{1-r_i^*}$, where α_i is independently sampled from a uniform distribution on $(0, 0.5]$ and k is a normalization coefficient chosen such that approximately 3% of samples have propensities below 0.01, creating a low-support tail. We then sample observations via $o_i \sim \text{Bernoulli}(p_i)$, using $\{(x_i, r_i^*)|o_i = 1\}$ for training and validation, while the original held-out test sets with complete labels serve as oracle evaluation sets.

Baselines. We compare the proposed **conRM** with the following baselines: (1) **classic debiasing methods**: IPS (Rosenbaum & Rubin, 1983; Wang et al., 2026a), DR (Robins et al., 1994; Wang et al., 2026a), CVIB (Wang et al., 2020), MTL-IPS (Zhang et al., 2020), and MTL-DR (Zhang et al., 2020); (2) **stabilized debiasing methods** for addressing small propensity scores: SNIPS (Swaminathan & Joachims, 2015), MRDR (Guo et al., 2021), DR-MSE (Dai et al., 2022), and StableDR (Li et al., 2022); the **Naive** method without any debiasing technique is included as the basic reference.

Experimental Details. We implement the reward model R_θ, propensity estimator p_ψ, and the imputation model e_ξ using an LLM backbone followed by MLP heads for all methods. To ensure fair comparison, the backbone is initialized with `FsfairX-LLaMA3-RM-v0.1`[1], and the MLP architecture is fixed to hidden dimensions of 256-64-1. We optimize the models using Adam (Kingma & Ba, 2014) for up to 600 epochs, employing early stopping with a patience of 30 epochs based on validation performance. Key hyperparameters are tuned on a validation set, with learning rate $\eta \in \{1 \times 10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}\}$, batch size $B \in \{512, 1024\}$, clipping threshold $\tau \in \{0.01, 0.05, 0.1\}$, nearest neighbors number $k \in \{10, 15, 20, 25\}$, and regularization coefficient $\beta \in \{0.01, 0.02, 0.05, 0.1, 0.2\}$. We report root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2) on the held-out test sets.

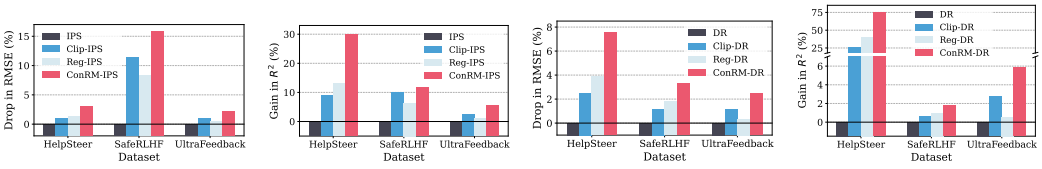
4.2 Performance Comparison

Table 1 summarizes the performance of different reward modeling approaches. The Naive method consistently performs worse than debiasing methods, demonstrating the necessity of correcting selection bias in observational feedback. IPS and DR substantially improve over the Naive baseline, while variance-reduction methods such as SNIPS and MRDR provide further gains by alleviating the impact of small propensities. However, these methods mainly focus on bias correction or weight

[1] https://huggingface.co/sfairXC/FsfairX-LLaMA3-RM-v0.1

Table 1: Performance comparison on three benchmark datasets. The best and second best results are in **bold** and underlined.

Method	HelpSteer			SafeRLHF			UltraFeedback		
	RMSE↓	R^2↑	MAE↓	RMSE↓	$R^{2\uparrow}$	MAE↓	RMSE↓	R^2↑	MAE↓
Naive	0.4575	0.0314	0.2535	0.2802	0.6837	0.1068	0.3811	0.3066	0.1897
CVIB	0.4502	0.0619	0.2527	0.2296	0.7878	0.1063	0.3489	0.4188	0.1840
IPS	0.4233	0.1708	0.2517	0.2665	0.7140	0.0983	0.3400	0.4481	0.1772
MTL-IPS	0.4296	0.1457	0.2498	0.2686	0.7094	0.0952	0.3334	0.4693	0.1774
SNIPS	0.4339	0.1482	0.2510	0.2328	0.7817	0.0977	0.3367	0.4589	0.1759
DR	0.4255	0.1620	0.2509	0.2305	0.7860	0.1020	0.3377	0.4556	0.1825
MTL-DR	0.4199	0.1840	0.2527	0.2249	0.7963	0.0954	0.3349	0.4648	0.1763
MRDR	0.4120	0.2144	0.2518	0.2269	0.7927	0.0968	0.3330	0.4707	0.1843
DR-MSE	0.4176	0.1930	0.2488	0.2281	0.7905	0.0978	0.3462	0.4280	0.1799
StableDR	0.4203	0.1629	0.2508	0.2308	0.7844	0.1013	0.3397	0.4491	0.1762
ConRM-IPS	0.4101	0.2218	0.2493	0.2244	0.7973	**0.0834**	0.3325	0.4724	**0.1680**
ConRM-DR	**0.3933**	**0.2842**	**0.2473**	**0.2229**	**0.7999**	0.0945	**0.3293**	**0.4823**	0.1740



(a) IPS with RMSE. (b) IPS with R^2. (c) DR with RMSE. (d) DR with R^2.

Figure 2: Ablation study of the ConRM framework under IPS and DR backbones across three benchmark datasets. The vertical axes show the relative decrease in RMSE and relative increase in R^2 compared with the corresponding debiasing baselines.

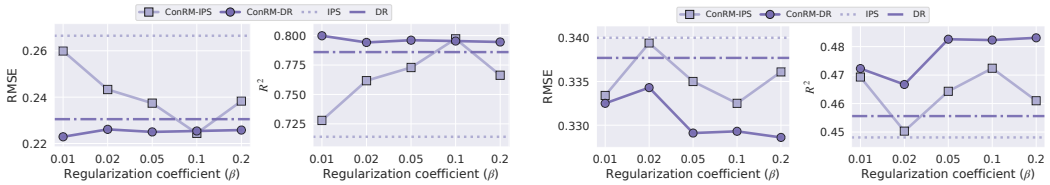
stabilization and do not explicitly constrain reward extrapolation in low-support regions. In contrast, ConRM consistently achieves the best performance across datasets, demonstrating that combining influence control with support-aware reward constraints is essential for reliable reward modeling under long-tailed observational feedback.

4.3 In-depth Analysis

Ablation study Figure 2 evaluates the contribution of different components in ConRM. Propensity clipping alone improves performance over the original IPS and DR objectives, confirming that limiting the influence of near-zero-propensity samples stabilizes reward learning. Conservative regularization further improves performance by constraining unreliable reward extrapolation in low-support regions. The full ConRM framework achieves the largest improvements across datasets, demonstrating that controlling sparse sample influence and preventing unsupported reward peaks are complementary for reliable observational reward modeling.

Effect of regularization coefficient β. Figure 3 studies the effect of the conservative regularization coefficient β. ConRM remains stable across different values, while moderate regularization generally achieves the best performance. This indicates that an appropriate conservative constraint can improve reward reliability without overly restricting model flexibility.

Generalization analysis Table 2 evaluates ConRM under different LLM backbones on PKU-SafeRLHF. Larger backbones improve the performance of all methods, showing the benefit of stronger representations. Nevertheless, ConRM consistently provides further gains over corresponding baselines under both Qwen3 and LLaMA2, indicating that its effectiveness is complementary to backbone scaling. The consistent advantage of ConRM-DR further highlights the robustness of combining doubly robust estimation with conservative reward constraints.



(a) Performance on SafeRLHF. (b) Performance on UltraFeedback.

Figure 3: Sensitivity analysis of the regularization coefficient β on SafeRLHF and UltraFeedback.

Table 2: Generalization analysis across different LLM backbones on PKU-SafeRLHF.

Method	RMSE↓	R^2↑	MAE↓
Qwen3-8B			
Naive	0.324	0.578	0.114
ConRM-IPS	0.297	0.643	0.097
ConRM-DR	**0.268**	**0.711**	**0.094**
LLaMA2-7B			
Naive	0.336	0.546	0.121
ConRM-IPS	0.307	0.620	0.108
ConRM-DR	**0.295**	**0.651**	**0.106**
Qwen3-14B			
Naive	0.286	0.669	0.090
ConRM-IPS	0.272	0.701	0.086
ConRM-DR	**0.263**	**0.721**	**0.084**
LLaMA2-13B			
Naive	0.303	0.630	0.112
ConRM-IPS	0.279	0.688	0.108
ConRM-DR	**0.268**	**0.711**	**0.103**

Table 3: Downstream RLHF safety performance. Δ denotes the relative improvement over the Naive baseline.

Benchmark	Method	Score↑	Δ (%)
HarmBench	Naive	0.794	–
	ConRM-IPS	0.833	5.0 ↑
	ConRM-DR	**0.861**	**8.6** ↑
JailbreakBench	Naive	0.700	–
	ConRM-IPS	**0.910**	**30.0** ↑
	ConRM-DR	0.890	27.1 ↑
StrongReject	Naive	0.796	–
	ConRM-IPS	0.882	10.8 ↑
	ConRM-DR	**0.890**	**11.8** ↑
WildGuard	Naive	0.660	–
	ConRM-IPS	0.736	11.5 ↑
	ConRM-DR	**0.752**	**14.0** ↑

4.4 Downstream RLHF Performance

Table 3 evaluates whether the improved reward reliability of ConRM translates into better downstream policy alignment. We use the reward models trained on PKU-SafeRLHF to fine-tune a Qwen2.5-7B policy with GRPO (Guo et al., 2025), and evaluate the resulting policies on HarmBench (Mazeika et al., 2024), JailbreakBench (Chao et al., 2024), StrongReject (Souly et al., 2024), and WildGuard (Han et al., 2024). Both ConRM-IPS and ConRM-DR consistently outperform the Naive baseline across all safety benchmarks, demonstrating that the benefits of ConRM extend beyond reward prediction accuracy to downstream policy optimization. These results further support our central motivation: constraining unsupported reward extrapolation provides a more reliable optimization signal and reduces the risk of policies exploiting spurious reward peaks.

5 Conclusion

Observational user feedback provides a scalable source of supervision for reward modeling, but its selective and long-tailed nature makes reliable learning challenging. In this work, we identify a critical gap between statistical unbiasedness and reward reliability: near-zero propensities can amplify sparsely supported observations into spurious reward peaks that may be exploited by downstream policy optimization. To address this issue, we propose Conservative Reward Modeling (ConRM), which combines propensity clipping with support-aware conservative reward constraints to limit sparse-sample influence and prevent unsupported reward extrapolation. Extensive experiments demonstrate that ConRM consistently improves reward estimation under long-tailed observational feedback and translates these gains into more reliable downstream RLHF alignment. One limitation is that the constructed reward intervals rely on local reward variations around well-supported anchors and may become unreliable when these patterns do not extend to low-support regions. Future work could explore more adaptive interval construction to reduce this dependence.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and improving the readability of the manuscript. They were not used to generate research ideas, design the methodology or experiments, formulate theoretical claims or proofs, interpret experimental results, or produce scientific conclusions. All AI-assisted text was carefully reviewed and revised by the authors. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Quanyu Dai, Haoxuan Li, Peng Wu, Zhenhua Dong, Xiao-Hua Zhou, Rui Zhang, Rui Zhang, and Jie Sun. A generalized doubly robust learning framework for debiasing post-click conversion rate prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 252–262, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Siyuan Guo, Lixin Zou, Yiding Liu, Wenwen Ye, Suqi Cheng, Shuaiqiang Wang, Hechang Chen, Dawei Yin, and Yi Chang. Enhanced doubly robust learning for debiasing post-click conversion rate estimation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275–284, 2021.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Advances in neural information processing systems*, 37:8093–8131, 2024.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Alex Qiu, Jiayi Zhou, Kaile Wang, Boxun Li, et al. Pku-saferlhf: Towards multi-level safety alignment for llms with human preference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31983–32016, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Haoxuan Li, Chunyuan Zheng, and Peng Wu. Stabledr: Stabilized doubly robust learning for recommendation on data missing not at random. *arXiv preprint arXiv:2205.04701*, 2022.

Yuhan Liu, Michael JQ Zhang, and Eunsol Choi. User feedback in human-llm dialogues: a lens to understand users but noisy as a learning signal. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2666–2681, 2025.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. Inform: Mitigating reward hacking in rlhf via information-theoretic reward modeling. *Advances in Neural Information Processing Systems*, 37:134387–134429, 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.

Maya L Petersen, Kristin E Porter, Susan Gruber, Yue Wang, and Mark J Van Der Laan. Diagnosing and responding to violations in the positivity assumption. *Statistical methods in medical research*, 21(1):31–54, 2012.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427):846–866, 1994.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *ICML*, 2016.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, et al. A strongreject for empty jailbreaks. *Advances in Neural Information Processing Systems*, 37:125416–125440, 2024.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *NeurIPS*, 2015.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hao Wang, Licheng Pan, Zhichao Chen, Chunyuan Zheng, Zhixuan Chu, Xiaoxi Li, Yuan Lu, Xinggao Liu, Haoxuan Li, and Zhouchen Lin. Causalrm: Causal-theoretic reward modeling for rlhf from observational user feedbacks. *arXiv preprint arXiv:2603.18736*, 2026a.

Hao Wang, Haocheng Yang, Licheng Pan, Lei Shen, Xiaoxi Li, Yinuo Wang, Zhichao Chen, Yuan Lu, Haoxuan Li, and Zhouchen Lin. Implicitrm: Unbiased reward modeling from implicit preference data for llm alignment. *arXiv preprint arXiv:2603.23184*, 2026b.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. Doubly robust joint learning for recommendation on data missing not at random. In *ICML*, 2019.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, et al. Helpsteer: Multi-attribute helpfulness dataset for steerlm. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3371–3384, 2024.

Zifeng Wang, Xi Chen, Rui Wen, Shao-Lun Huang, Ercan Kuruoglu, and Yefeng Zheng. Information theoretic counterfactual learning from missing-not-at-random feedback. *Advances in Neural Information Processing Systems*, 33:1854–1864, 2020.

Wenhao Zhang, Wentian Bao, Xiao-Yang Liu, Keping Yang, Quan Lin, Hong Wen, and Ramin Ramezani. Large-scale causal approaches to debiasing post-click conversion rate estimation with multi-task learning. In *Proceedings of the web conference 2020*, pp. 2775–2781, 2020.

A Appendix

A.1 Learning Algorithm

Algorithm 1 summarizes ConRM with either an IPS or DR debiasing backbone.

Algorithm 1: Optimization Workflow of ConRM

Input: Population $\mathcal{D}$ with indicators o_i and labels r_i when $o_i = 1$; propensity estimates $\hat{p}_i$; backbone $b \in \{\mathrm{IPS}, \mathrm{DR}\}$; threshold τ; regularization weight β; neighbors k; learning rates η_θ, η_ξ.

Output: Reward model parameters θ.

Partition $\mathcal{D}$ into $\mathcal{D}_{\mathrm{well}}$ and $\mathcal{D}_{\mathrm{low}}$ using Eq. equation 14;
while *the stopping criterion is not met* **do**
  Sample a mini-batch $\mathcal{B} \subseteq \mathcal{D}$ and set $\mathcal{B}_{\mathrm{low}} \leftarrow \mathcal{B} \cap \mathcal{D}_{\mathrm{low}}$;
  Compute $\tilde{p}_i \leftarrow \max(\hat{p}_i, \tau)$ for $x_i \in \mathcal{B}$ using Eq. equation 10;
  `// Influence-controlled debiasing`
  if $b = \mathrm{DR}$ **then**
    Update ξ using Eq. equation 8 on $\mathcal{B}$ with $\hat{p}_i$ replaced by $\tilde{p}_i$;
    Compute $\mathcal{L}_{\mathrm{debias}} \leftarrow \mathcal{L}_{\mathrm{DR}}^{\mathrm{clip}}$ using Eq. equation 12;
  else
    Compute $\mathcal{L}_{\mathrm{debias}} \leftarrow \mathcal{L}_{\mathrm{IPS}}^{\mathrm{clip}}$ using Eq. equation 11;
  end
  `// Support-aware conservative constraint`
  foreach $x \in \mathcal{B}_{\mathrm{low}}$ **do**
    Retrieve $z(x)$ from $\mathcal{D}_{\mathrm{well}}$ and compute $v(x)$ using Eqs. equation 16 and equation 17;
    Retrieve $\mathcal{N}_k(z(x)) \subseteq \mathcal{D}_{\mathrm{well}}$ and compute $\hat{D}_{\min}(x), \hat{D}_{\max}(x)$ using Eqs. equation 18 and equation 19;
    Construct $R_{\mathrm{lower}}(x), R_{\mathrm{upper}}(x)$ using Eqs. equation 20 and equation 21;
  end
  Compute the mini-batch regularizer $\mathcal{R}_{\mathrm{con}}$ using Eq. equation 22; set $\mathcal{R}_{\mathrm{con}} = 0$ if $\mathcal{B}_{\mathrm{low}} = \varnothing$;
  Compute $\mathcal{L}_{\mathrm{ConRM}} \leftarrow \mathcal{L}_{\mathrm{debias}} + \beta \mathcal{R}_{\mathrm{con}}$ using Eq. equation 9;
  Update $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{\mathrm{ConRM}}$;
end

A.2 Proofs

Proposition 1 (Coverage of conservative reward intervals) *For any $x \in \mathcal{D}_{\mathrm{low}}$, if Assumption 1 holds, the true reward value $R^*(x)$ is covered by the constructed interval up to estimation errors:*

$$R^*(x) \in [R_{\mathrm{lower}}(x) - \epsilon(x), R_{\mathrm{upper}}(x) + \epsilon(x)],$$

where $\epsilon(x)$ captures the estimation errors of the anchor reward and the local reward variation. When these quantities are estimated exactly, $\epsilon(x) = 0$.

Proof. For a low-support sample $x \in \mathcal{D}_{\mathrm{low}}$, let $z(x)$ denote its well-supported anchor and define

$$d = \|\phi(x) - \phi(z(x))\|_2,$$

where $\phi(\cdot)$ denotes the representation in the hidden space. We further define the unit direction vector from the anchor to x as

$$v(x) = \frac{\phi(x) - \phi(z(x))}{d}.$$

The straight path between the anchor and the target sample can be parameterized as

$$\gamma(t) = \phi(z(x)) + tv(x), \quad t \in [0, d].$$

By the chain rule and the fundamental theorem of calculus, the true reward at x can be expressed as

$$R^*(x) = R^*(z(x)) + \int_0^d \nabla_\phi R^*(\gamma(t))^\top v(x)\, dt.$$

The term $\nabla_\phi R^*(\gamma(t))^\top v(x)$ represents the directional reward variation along the path from the anchor to the low-support sample.

According to Assumption 1, the reward variation along this path is bounded by the empirical variation range observed around neighboring well-supported samples. Therefore, there exist non-negative estimation errors $\delta_{\min}(t)$ and $\delta_{\max}(t)$ such that

$$\hat{D}_{\min}(x;\theta) - \delta_{\min}(t) \le \nabla_\phi R^*(\gamma(t))^\top v(x) \le \hat{D}_{\max}(x;\theta) + \delta_{\max}(t).$$

Integrating both sides over $t \in [0, d]$ gives

$$\begin{aligned} d\hat{D}_{\min}(x;\theta) - \Delta_{\min}(x) &\le \int_0^d \nabla_\phi R^*(\gamma(t))^\top v(x)\, dt \\ &\le d\hat{D}_{\max}(x;\theta) + \Delta_{\max}(x), \end{aligned}$$

where $\Delta_{\min}(x) = \int_0^d \delta_{\min}(t)dt$ and $\Delta_{\max}(x) = \int_0^d \delta_{\max}(t)dt$ denote the accumulated directional variation estimation errors.

Substituting this bound into the line integral expression yields

$$\begin{aligned} R^*(x) &\ge R^*(z(x)) + d\hat{D}_{\min}(x;\theta) - \Delta_{\min}(x), \\ R^*(x) &\le R^*(z(x)) + d\hat{D}_{\max}(x;\theta) + \Delta_{\max}(x). \end{aligned}$$

Let $\epsilon_R(z(x))$ denote the anchor reward estimation error:

$$R^*(z(x)) = R_\theta(z(x)) + \epsilon_R(z(x)).$$

Then,

$$\begin{aligned} R^*(x) &\ge R_\theta(z(x)) + d\hat{D}_{\min}(x;\theta) + \epsilon_R(z(x)) - \Delta_{\min}(x), \\ R^*(x) &\le R_\theta(z(x)) + d\hat{D}_{\max}(x;\theta) + \epsilon_R(z(x)) + \Delta_{\max}(x). \end{aligned}$$

Recall that the constructed interval boundaries are

$$\begin{aligned} \hat{R}_{\text{lower}}(x) &= R_\theta(z(x)) + d\hat{D}_{\min}(x;\theta), \\ \hat{R}_{\text{upper}}(x) &= R_\theta(z(x)) + d\hat{D}_{\max}(x;\theta). \end{aligned}$$

Therefore,

$$\begin{aligned} R^*(x) &\ge \hat{R}_{\text{lower}}(x) + \epsilon_R(z(x)) - \Delta_{\min}(x), \\ R^*(x) &\le \hat{R}_{\text{upper}}(x) + \epsilon_R(z(x)) + \Delta_{\max}(x). \end{aligned}$$

By defining

$$\epsilon(x) = |\epsilon_R(z(x))| + \max\{\Delta_{\min}(x), \Delta_{\max}(x)\},$$

we obtain the final interval guarantee:

$$R^*(x) \in [\hat{R}_{\text{lower}}(x) - \epsilon(x), \hat{R}_{\text{upper}}(x) + \epsilon(x)].$$

When the anchor reward is estimated exactly and the empirical directional variation perfectly captures the true reward variation, we have

$$\epsilon_R(z(x)) = 0, \quad \Delta_{\min}(x) = \Delta_{\max}(x) = 0,$$

which gives $\epsilon(x) = 0$. □