Diffusion Reward Models

Anonymous ACL submission

Abstract

Reward models underpin the alignment of large language models, yet the dominant designs reduce each prompt–response pair to a point estimate or to a distribution from a fixed parametric family. This is at odds with human preference, which is inherently multimodal: the same response can be reasonably judged in many ways, and no single family covers all of them. To better fit this structure, we introduce **DRM**, a **D**iffusion **R**eward **M**odel that recasts reward modeling as conditional density estimation over $p(\mathbf{r} \mid x, y)$. Conditioned on a frozen LLM encoder, a lightweight Diffusion Transformer denoises Gaussian noise into a reward vector, placing no parametric assumption on the output distribution and naturally representing its multimodal structure. A single architecture handles both multi-attribute regression and pairwise preference data, and at inference N samples form an empirical reward distribution that can be aggregated into a scalar, a variance, or quantiles. Across five benchmarks, DRM matches or surpasses baselines under matched data and backbone, stays competitive with much larger discriminative, distributional, and generative RMs despite its modest training scale, and recovers multimodal reward structure where conventional heads collapse to a point.

1 Introduction

Reward models (RMs) are critical to post-training for large language models (LLMs). In Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022), the RM defines the optimization signal, and its quality directly determines whether alignment improves or degenerates into reward hacking (Gao et al., 2023). While verifiable domains such as mathematics and code admit ground-truth signals that enable RLVR-style training without a learned reward (Lambert et al., 2024; Guo et al., 2025), general-domain alignment has no such oracle and remains dependent on learned RMs. The dominant paradigms in this context are discriminative RMs trained with Bradley–Terry (BT) losses (Bradley and Terry, 1952) and generative RMs trained with next-token prediction (Mahan et al., 2024; Zhang et al., 2025). Both ultimately collapse the model's output into a deterministic scalar score $r(x, y) \in \mathbb{R}$. This assumes that for any prompt x and response y there exists a stable point-valued reward.

This assumption is at odds with how human preference behaves: it is **multimodal**[1]. Annotators disagree systematically over values, rubric interpretation, and helpfulness–harmlessness tradeoffs, with inter-annotator agreement on Anthropic-HH only $\sim$63% (Bai et al., 2022) and substantial within-rubric disagreement persists in HelpSteer3-Preference even after careful filtering (Wang et al., 2026). Siththaranjan et al. (2024) shows that BT training on data with hidden context implicitly applies a Borda-count rule, diverging from risk-neutral expected utility, while multi-objective works (Wang et al., 2024a,d,c) show that a single scalar is a lossy projection of a multi-attribute reward vector. Collapsing $p(r \mid x, y)$ to a point estimate erases this disagreement, uncertainty, and multimodal structure.

Existing attempts to escape the scalar bottleneck remain partial: each commits to a specific output-distribution family that cannot represent multimodal reward distributions. Multi-objective RMs (Wang et al., 2024a,c) require a fixed, predefined per-attribute schema and produce a vector that is then collapsed by a gate. Generative and rubric-based judges (Guo et al., 2026; Chen et al., 2025b; Gunjal et al., 2025; Viswanathan

[1] Multimodal is used here in its statistical sense: a distribution with multiple modes or local peaks; see https://en.wikipedia.org/wiki/Multimodal_distribution. It does not refer to the common usage of multiple input/output modalities such as text, image, or audio.

et al., 2026) shift modeling burden to long-form reasoning at steep inference cost, yet still output a single verdict. Parametric distributional heads commit explicitly: DPL (Siththaranjan et al., 2024) and URM (Lou et al., 2024) predict a Gaussian mean–variance and are unimodal by construction; DPRM (Li et al., 2024) predicts a categorical distribution and is limited by bin granularity; QRM (Dorka, 2024) predicts a fixed grid of quantiles and suffers from quantile crossing and unstable tails. What is missing is a reward head that does not commit to an output family and is expressive enough to represent the multimodal distribution that human preference actually induces.

We propose **DRM**, a **D**iffusion **R**eward **M**odel that replaces the conventional value head with a diffusion-based head. Conditioned on a frozen LLM backbone's last-token hidden state, a lightweight Diffusion Transformer (DiT) (Peebles and Xie, 2023) denoises Gaussian noise into a K-dimensional reward vector, directly modeling $p(\mathbf{r} \mid x, y)$. Unlike heads that assume a parametric output family, a Diffusion Reward Head places no such constraint on the distribution it represents and is known for strong multimodal coverage where alternatives collapse or blur (Dhariwal and Nichol, 2021; Song et al., 2020b), making it a natural fit for the multimodal rewards human preference induces. A single architecture then serves both data regimes: with $K > 1$ it fuses heterogeneously-labeled multi-attribute data, and with $K = 1$ a distributional BT objective trains it directly on preference pairs. At inference, N samples from the head form an empirical reward distribution that can be aggregated into a scalar for standard RLHF, a variance for uncertainty, or quantiles for risk-averse selection.

We evaluate DRM on a frozen LLM encoder across five standard RM benchmarks: RewardBench v2 (Malik et al., 2025), PPE (Frick et al., 2025), RMB (Zhou et al., 2025), RM-Bench (Liu et al., 2025b), and JudgeBench (Tan et al., 2025), spanning chat, instruction following, math, code, factuality, and safety. Trained on identical data and backbone, DRM matches or surpasses the scalar head ArmoRM (Wang et al., 2024a) and the parametric-quantile head QRM (Dorka, 2024), is competitive with strong discriminative RMs, and approaches generative judges such as GPT-4o at far lower inference cost. We further visualize the learned $p(\mathbf{r} \mid x, y)$ on examples with divergent preference signals, showing that DRM recovers multimodal reward structure where other RMs collapse to a point.

We summarize our contributions as follows:

- We identify a shared limitation across scalar, multi-attribute, and parametric-distributional RMs: all commit to a fixed output-distribution family. We instead recast reward modeling as density estimation over $p(\mathbf{r} \mid x, y)$.
- We introduce **DRM**, a new reward-modeling paradigm that replaces the value head with a DiT head imposing no parametric form on the output, making it well-suited to the multimodal reward distributions that human preference induces. A single architecture covers both multi-attribute regression and a distributional BT objective on preference pairs, and the head opens a reward-axis test-time scaling absent from current RMs.
- As an initial empirical validation across five RM benchmarks, DRM matches or surpasses baselines under matched data and backbone, and stays competitive with much larger discriminative, distributional, and generative RMs despite its modest training scale, while recovering multimodal reward distributions where conventional RMs collapse to a point.

2 Method

DRM replaces the deterministic scalar value head of conventional reward models with a Diffusion Reward Head that directly models $p(\mathbf{r} \mid x, y)$ without committing to any parametric family. As shown in Figure 1, a frozen LLM encoder produces a representation $\mathbf{h}$ of (x, y), a lightweight Diffusion Transformer (DiT) denoises Gaussian noise into reward vectors conditioned on $\mathbf{h}$ (Section 2.1), trained either on multi-attribute or pairwise preference data (Section 2.2), and N samples form an empirical distribution at inference (Section 2.3).

2.1 Architecture

We explore a complementary direction to existing reward modeling methods: we model $p(\mathbf{r} \mid x, y)$ implicitly through an iterative denoising process. This is motivated by the well-documented mode coverage of diffusion models (Dhariwal and Nichol, 2021) and their ability to approximate arbitrary continuous densities (Song et al., 2020b; Lipman et al., 2022), which makes them well-suited to the multimodal reward distributions induced by human preference. Concretely, DRM consists of two components: a frozen LLM encoder and a Diffusion Reward Head, followed by a non-parametric aggregator described in Section 2.3.

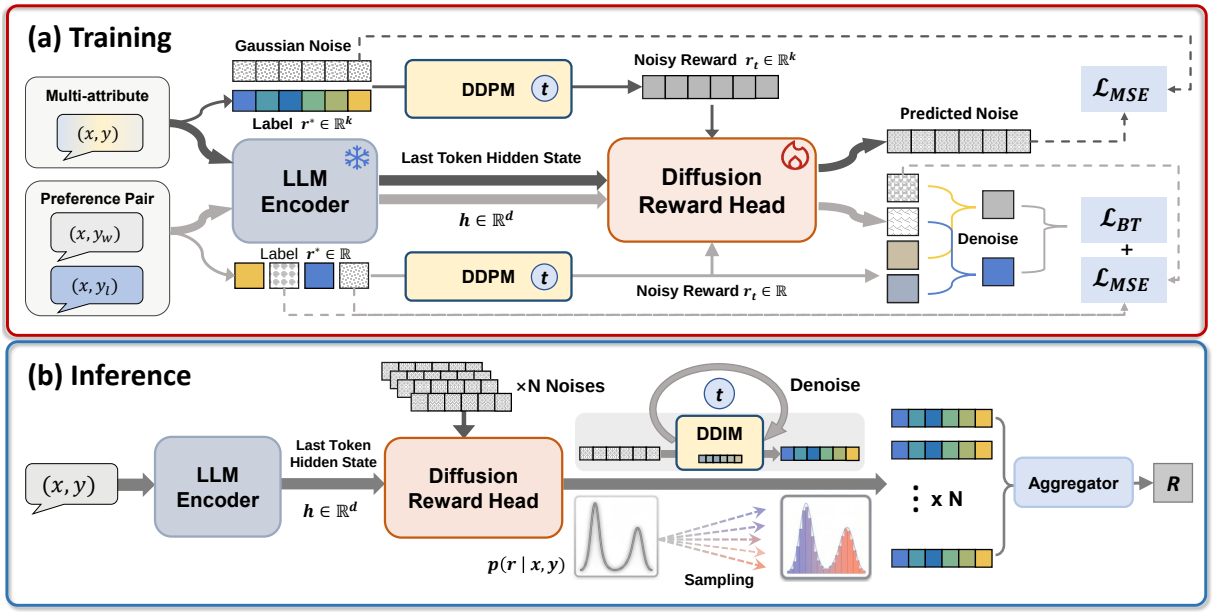

Figure 1: Overview of DRM. A frozen encoder maps (x, y) to a hidden state $\mathbf{h}$ that conditions a Diffusion Reward Head modeling $p(\mathbf{r} \mid x, y)$. **(a) Training.** A single head supports both multi-attribute regression, supervised by a masked denoising loss $\mathcal{L}_{\text{MSE}}$, and pairwise preference data, supervised by an additional Bradley–Terry objective $\mathcal{L}_{\text{BT}}$ on the denoised rewards. **(b) Inference.** Drawing N samples via DDIM yields an empirical reward distribution that captures multimodal structure and is summarized by an aggregator into the final reward R.

LLM Encoder. For each prompt–response pair (x, y), we use a frozen language model as the backbone encoder. We format the prompt and response into a single input sequence, feed it into the encoder, and take the hidden state of the last token as the semantic representation of the sample:

$$\mathbf{h} = \text{Enc}(x, y) \in \mathbb{R}^{d_{\text{enc}}}.$$

In our implementation, this encoding process is performed offline: we precompute $\mathbf{h}$ for each (x, y), and then train the Diffusion Reward Head on top of these frozen representations. This decouples costly language representation learning from subsequent reward distribution modeling, preserving the semantic priors of the pretrained encoder while substantially reducing training cost.

Diffusion Reward Head. The DiT reward head ϵ_θ is a lightweight DiT that models the conditional distribution of a K-dimensional reward vector:

$$p_\theta(\mathbf{r} \mid \text{Enc}(x, y)) = p_\theta(\mathbf{r} \mid \mathbf{h}), \qquad \mathbf{r} \in \mathbb{R}^K.$$

Given a noisy reward $\mathbf{r}_t$ and timestep t, the head predicts the noise during the forward process,

$$\hat{\boldsymbol{\epsilon}} = \epsilon_\theta(\mathbf{r}_t, t, \mathbf{h}),$$

where the textual condition $\mathbf{h}$ and timestep t are injected into each DiT block via adaptive layer normalization (Peebles and Xie, 2023). Full implementation details including projection layers, sinusoidal timestep embedding, the conditioning fusion, and initialization are deferred to Appendix A.

2.2 Training

DRM is trained within a unified diffusion framework that supports both multi-attribute regression and pairwise preference data. Both regimes share the same diffusion parameterization and differ only in how supervision targets are constructed.

Probabilistic Formulation. Our goal is to learn the conditional distribution

$$p_\theta(\mathbf{r} \mid x, y, \mathbf{m}), \qquad \mathbf{r} \in \mathbb{R}^K,$$

where the reward space has dimension $K = 1$ for scalar reward and $K > 1$ for multi-attribute reward vectors, and $\mathbf{m} \in \{0, 1\}^K$ indicates which dimensions are annotated for (x, y). The mask is introduced so that a single model can be trained on multi-attribute datasets within a unified reward space, with unlabeled dimensions excluded from supervision. Since $\mathbf{m}$ is data-specific, we omit it from the notation where it is clear from context.

Training Objective. We define two complementary objectives sharing the same noise-prediction parameterization.

Multi-attribute regression. For data with explicit reward annotations $\mathbf{r}_0$, we train the head with a

Algorithm 1 Training of DRM

Require: Frozen backbone Enc, DiT reward head ϵ_θ, training set $\mathcal{D}$, noise schedule $\{\bar{\alpha}_t\}_{t=1}^{T}$, CFG rate p_{drop}

for each training iteration **do**
  Sample a minibatch $\mathcal{B} \subset \mathcal{D}$
  for each sample $(x, y) \in \mathcal{B}$ **do**
    Obtain semantic embedding $\mathbf{h} = \text{Enc}(x, y)$
    Construct reward target $\mathbf{r}_0$ and mask $\mathbf{m}$
    With probability , replace $\mathbf{h}$ with the unconditional vector
    Sample $t \sim \text{U}(\{1, \ldots, T\})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    Form $\mathbf{r}_t = \sqrt{\bar{\alpha}_t}\mathbf{r}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}$
    Predict $\hat{\boldsymbol{\epsilon}} = \epsilon_\theta(\mathbf{r}_t, t, \mathbf{h}, \mathbf{m})$
  end for
  if $\mathcal{D}$ is multi-attribute **then**
    Compute $\mathcal{L}_{\text{denoise}}$ over the batch
  else ▷ pairwise preference
    Recover $\hat{\mathbf{r}}_0 = (\mathbf{r}_t - \sqrt{1 - \bar{\alpha}_t}\hat{\boldsymbol{\epsilon}})/\sqrt{\bar{\alpha}_t}$ for each response
    Compute $\mathcal{L}_{\text{pair}} = \mathcal{L}_{\text{denoise}} + \lambda_{\text{BT}}\mathcal{L}_{\text{BT}}$
  end if
  Update ϵ_θ
end for

Algorithm 2 Inference of DRM

Require: Prompt-response pair (x, y), encoder Enc, DiT reward head ϵ_θ, DDIM scheduler, sampling steps S, number of reward samples N, guidance scale ω, aggregation function $\mathcal{A}(\cdot)$

Compute the semantic embedding $\mathbf{h} = \text{Enc}(x, y)$
for $i = 1$ to N **do**
  Sample initial Gaussian noise $\mathbf{r}_T^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  for t in DDIM timesteps **do**
    Predict unconditional noise $\hat{\boldsymbol{\epsilon}}_{\text{uncond}} = \epsilon_\theta(\mathbf{r}_t^{(i)}, t, \varnothing)$
    Predict conditional noise $\hat{\boldsymbol{\epsilon}}_{\text{cond}} = \epsilon_\theta(\mathbf{r}_t^{(i)}, t, \mathbf{h})$
    Apply classifier-free guidance $\hat{\boldsymbol{\epsilon}} = \hat{\boldsymbol{\epsilon}}_{\text{uncond}} + \omega\,(\hat{\boldsymbol{\epsilon}}_{\text{cond}} - \hat{\boldsymbol{\epsilon}}_{\text{uncond}})$
    Perform one DDIM reverse step $\mathbf{r}_{t-1}^{(i)} = \text{DDIMStep}(\mathbf{r}_t^{(i)}, \hat{\boldsymbol{\epsilon}}, t)$
  end for
  Obtain one reward sample $\mathbf{r}^{(i)} \leftarrow \mathbf{r}_0^{(i)}$
end for
Form the empirical reward distribution $\hat{P}(\mathbf{r} \mid x, y)$
return $Agg\Big(\{\mathbf{r}^{(i)}\}_{i=1}^{N}\Big)$ ▷ Aggregation

masked denoising loss that minimizes the mean squared error between the predicted and ground-truth noise over the annotated dimensions only:

$$\mathcal{L}_{\text{denoise}} = \frac{1}{\|\mathbf{m}\|_1}\Big\|\mathbf{m} \odot \big(\epsilon_\theta(\mathbf{r}_t, t, \mathbf{h}, \mathbf{m}) - \boldsymbol{\epsilon}\big)\Big\|_2^2,$$

where the forward process is $\mathbf{r}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{r}_0 + \sqrt{1 - \bar{\alpha}_t}\,\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Under Gaussian diffusion, this objective is equivalent to learning the score function of the conditional reward distribution (Ho et al., 2020; Song et al., 2020b).

Pairwise preference. For data containing only pairwise preference, pointwise regression is insufficient to capture which response is preferred. For each pair (x, y_w, y_l), let $\hat{r}_0^{(w)}$ and $\hat{r}_0^{(l)}$ denote the denoised reward estimates obtained from the standard reconstruction $\hat{\mathbf{r}}_0 = (\mathbf{r}_t - \sqrt{1 - \bar{\alpha}_t}\,\hat{\boldsymbol{\epsilon}})/\sqrt{\bar{\alpha}_t}$. We impose a Bradley–Terry-style ranking objective on the denoised estimates,

$$\mathcal{L}_{\text{BT}} = -\log \sigma\Big(\hat{r}_0^{(w)} - \hat{r}_0^{(l)}\Big).$$

To preserve the denoising signal alongside the ranking signal, we additionally apply $\mathcal{L}_{\text{denoise}}$ on them. The final pairwise loss is

$$\mathcal{L}_{\text{pair}} = \mathcal{L}_{\text{denoise}} + \lambda_{\text{BT}}\mathcal{L}_{\text{BT}}.$$

Training Procedure. The full training procedure is summarized in Algorithm 1 and visualized in Figure 1 (a). During training, we additionally apply classifier-free guidance (CFG) (Ho and Salimans, 2022). With a fixed probability, the textual condition $\mathbf{h}$ is replaced by a learnable unconditional vector, which enables guided sampling at inference time (Section 2.3). Training datasets and hyperparameter settings are reported in Section 3.1.

2.3 Inference

At inference time, we first estimate the conditional reward distribution by sampling from the trained Diffusion Reward Head, and then summarize it into the statistic required by the downstream task. Unlike conventional reward models that only produce a single scalar, the inference process of DRM naturally preserves distribution-level reward information. Therefore, DRM can be used not only for standard reward scoring, but also for test-time scaling based on distribution.

Reward Distribution Sampling. Given (x, y) and its semantic embedding $\mathbf{h}$, we draw N independent samples from the reward head by running the reverse process from standard Gaussian noise,

$$\mathbf{r}^{(1)}, \ldots, \mathbf{r}^{(N)} \sim p_\theta(\mathbf{r} \mid \mathbf{h}),$$

which together form an empirical approximation to the reward distribution. We use DDIM (Song et al., 2020a) to reduce the number of reverse steps and combine it with CFG controlled by a guidance scale ω; the full procedure is summarized in Algorithm 2 and visualized in Figure 1 (b).

Aggregations. After obtaining the empirical distribution, we can exploit this distributional information in ways that distinguish it from conven-

tional scalar reward models. In this work we use the **mean** $\bar{\mathbf{r}} = \frac{1}{N}\sum_i \mathbf{r}^{(i)}$ averaged across reward dimensions for an empirical study, which yields a scalar reward compatible with standard RLHF, Best-of-N selection, and reward-benchmark protocols. The same samples also support distributional summaries unavailable to scalar RMs, such as the **variance** as an uncertainty or anti-reward-hacking signal, and **quantiles** or the empirical CDF for risk-sensitive decision rules; we leave a systematic study of these to future work.

Two Axes of Test-Time Scaling. DRM exposes two independent test-time scaling axes. Along the *response axis*, the classical Best-of-N approach generates N candidate responses and selects the one with the highest aggregated reward, $y^\star = \arg\max_{y_i} \bar{r}(x, y_i)$, which is shared with any scalar reward model. Along the *reward axis*, increasing the number of diffusion samples N for a fixed (x, y) tightens the estimate of $\bar{\mathbf{r}}(x, y)$, improving scoring precision without changing the candidate set. **The second axis is unique to DRM: a deterministic scalar RM produces the same score for any N, so no analogous lever exists.** We empirically study both axes in Section 3.3.

3 Experiments

In this section, we evaluate DRM's overall performance and test-time scaling capability, followed by fine-grained analyses of its behavior and its ability to capture multimodal reward distributions.

3.1 Experimental Setup

We train two DRM variants under the two supervision regimes of Section 2.2: **DRM-Multi-8B**, trained on multi-attribute reward data, and **DRM-Pref-8B**, trained on pairwise preference data. Both DRM variants use the LLM encoder of FsfairX-LLaMA3-RM-v0.1 (Dong et al., 2023) as a frozen backbone, with its scalar value head removed and replaced by our Diffusion Reward Head.

Training data. DRM-Multi-8B is trained on the aggregated multi-attribute reward corpus of ArmoRM (Wang et al., 2024a), which unifies several public reward-modeling sources into 569K samples annotated over 19 attributes (helpfulness, coherence, instruction following, etc.), each example labeling only a subset. DRM-Pref-8B is trained on the Tulu3 preference mixture (Lambert et al., 2024) (273K chosen/rejected pairs, no explicit reward labels). The two regimes follow the masked denoising and distributional Bradley–Terry objectives of Section 2.2 respectively.

	DRM-Multi-8B	**DRM-Pref-8B**
Training Data	ArmoRM agg.	Tulu3 pref.
Reward Dim K	19	1
Diffusion Reward Head		
Hidden Size	384	384
Blocks / Heads	3 / 6	3 / 6
Dropout	0.2	0.2
Training		
Max Diffusion Steps	1000	1000
Beta Schedule	sqcos	sqcos
Batch Size	64	128
Learning Rate	5×10^{-5}	3×10^{-5}
λ_{BT}	–	0.5
Inference		
DDIM Steps	10	10
Guidance ω	7	7
# samples N	32	32

Table 1: Configurations of the two DRM variants.

Benchmarks. We evaluate on five public reward-model benchmarks: RewardBench v2 (Malik et al., 2025), PPE (Frick et al., 2025), RMB (Zhou et al., 2025), RM-Bench (Liu et al., 2025b), and JudgeBench (Tan et al., 2025). Together they span chat, instruction following, mathematics, code, reasoning, factuality, and safety, and probe complementary protocols like pairwise accuracy and Best-of-N selection. Per-benchmark statistics and task compositions are deferred to Appendix B.

Baselines. We compare against three families of reward models. **Discriminative RMs** directly output a deterministic scalar score (e.g., ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024a), Skywork-Reward-Llama-3.1-8B-v0.2 (Liu et al., 2024) and InternLM2-20B-Reward (Cai et al., 2024)). **Generative RMs** emit a textual verdict or reasoning trace before a score (e.g., DeepSeek-GRM-27B (Liu et al., 2025c), the closed-source GPT-4o (OpenAI et al., 2024) and Claude-3.5-Sonnet (Anthropic, 2024)). **Parametric distributional RMs** model the reward distribution within a fixed family (QRM-Gemma-2-27B (Dorka, 2024), URM-Llama-3.1-8B (Lou et al., 2024)). DRM differs from all three by modeling $p(\mathbf{r} \mid x, y)$ non-parametrically, which additionally yields uncertainty and distributional statistics rather than a point estimate. These comparisons let us ask whether explicit distributional modeling helps on standard reward-modeling tasks, and whether a diffusion head is more expressive than existing parametric-distributional approaches.

DRM configuration. Both variants share the same architecture and diffusion schedule, differing only

Table 2: Overall benchmark results of different reward models. Baselines are taken from Liu et al. (2025a) where available; entries not reported therein are evaluated by us under the same protocol. **Gray rows denote large-scale or closed baselines that are not directly comparable in model size or publicly reported training-data scale.**

Reward Models	RewardBench v2	PPE Pref	PPE Corr	RMB Pairwise	RM-Bench	JudgeBench	Avg.
Discriminative Reward Models							
internlm2-7b-reward (Cai et al., 2024)	53.4	62.1	60.4	67.1	67.1	59.4	61.6
Eurus-RM-7b (Yuan et al., 2024)	58.1	59.6	60.0	65.5	69.0	58.4	61.8
Starling-RM-34B (Zhu et al., 2023)	45.5	62.8	60.3	72.0	67.1	63.8	61.9
RM-Mistral-7B (Dong et al., 2023)	59.6	61.8	56.4	66.6	66.9	62.1	62.2
ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024a)	66.5	60.6	61.4	64.6	67.7	53.2	62.3
internlm2-20b-reward (Cai et al., 2024)	56.3	61.0	63.0	62.9	72.1	64.3	63.3
Llama-3-OffsetBias-RM-8B (Park et al., 2024)	64.8	59.2	64.1	57.8	71.3	63.5	63.5
FsfairX-LLaMA3-RM-v0.1 (Dong et al., 2023)	62.9	63.1	61.1	70.2	71.7	56.6	64.3
Skywork-Reward-Llama-3.1-8B-v0.2 (Liu et al., 2024)	71.8	62.2	60.7	66.6	64.7	62.9	64.8
GRM-Llama3-8B-rewardmodel-ft (Yang et al., 2024)	67.7	62.1	60.0	70.2	69.9	62.3	65.4
Llama-3.1-Nemotron-70B-Reward (Wang et al., 2024b)	76.7	64.2	63.2	64.9	72.2	65.8	67.8
Skywork-Reward-V2-Qwen3-8B (Liu et al., 2025a)	78.2	70.6	75.1	81.2	82.6	73.4	76.9
Generative Reward Models							
DeepSeek-GRM-27B (Liu et al., 2025c)	64.4	64.7	59.8	69.0	72.4	63.0	65.6
GPT-4o (OpenAI et al., 2024)	64.9	67.7	67.1	73.8	73.1	59.8	67.7
Claude-3.5-Sonnet (Anthropic, 2024)	64.7	67.3	69.2	70.6	74.5	64.8	68.5
Parametric Distributional Reward Models							
QRM-Gemma-2-27B (Dorka, 2024)	76.7	52.3	54.8	53.4	65.9	57.5	60.1
QRM-Llama3.1-8B-v2 (Dorka, 2024)	70.7	57.2	60.3	61.1	72.5	62.6	64.1
URM-LLaMa-3.1-8B (Lou et al., 2024)	73.9	60.2	60.4	65.7	72.0	64.1	66.1
LDL-Reward-Gemma-2-27B-v0.1 (Chen et al., 2025a)	72.5	62.4	63.9	67.9	71.0	64.2	67.0
Diffusion Reward Models (Ours)							
DRM-Multi-8B	65.6	62.5	63.8	78.0	68.8	58.6	66.2
DRM-Pref-8B	65.7	63.0	62.5	78.2	68.1	57.1	65.8

in reward dimension K and a few optimization settings. Full configurations are listed in Table 1. All hyperparameters are tuned on RewardBench v2 and then fixed across the remaining benchmarks. The hyperparameter search is detailed in Appendix C.

3.2 Main Results

Under matched data and backbone, the diffusion head wins. Table 2 reports results across the five benchmarks. We position DRM as a first exploration of a native distributional reward head, so the most controlled comparison is against models trained on the same ArmoRM corpus with the same backbone, where only the reward head differs. Here DRM-Multi-8B reaches an average of 66.2, exceeding both the scalar multi-attribute head ArmoRM (62.3) and the parametric-quantile head QRM (64.1); the gain is largest on RMB (78.2). With data and backbone held fixed, this isolates the reward head as the source of the improvement.

DRM stays competitive beyond the matched setting. Beyond this controlled comparison, DRM remains competitive with strong discriminative RMs trained on other data or larger backbones, such as Eurus-RM-7B (61.8), Skywork-Reward-Llama-3.1-8B-v0.2 (64.8), and Llama-3.1-Nemotron-70B (67.8), and approaches or even surpasses generative judges that rely on far larger capacity and explicit reasoning at much higher cost, e.g. DeepSeek-GRM-27B (65.6) and GPT-4o (67.7). Despite its modest training scale, a non-parametric diffusion head is thus competitive across discriminative, distributional, and generative families.

A non-parametric head matches parametric ones while staying more general. The parametric distributional models share DRM's goal of modeling $p(\mathbf{r} \mid x, y)$, but realize it with an MLP that emits the parameters of a fixed family in a single forward pass, like a Gaussian for URM or a fixed quantile grid for QRM. DRM instead represents the distribution implicitly through iterative denoising, with no parametric form imposed on its shape. It performs on par with these models on average (66.2, comparable to URM's 66.1 and QRM's 64.1) while making no distributional assumption, suggesting diffusion reward modeling offers a flexible and viable alternative.

The framework transfers across supervision types. Finally, DRM-Multi-8B and DRM-Pref-8B perform similarly (66.2 vs. 65.8) despite being trained under multi-attribute regression versus pairwise preference, indicating that the same reward-space diffusion framework transfers across both regimes. We defer a discussion of remaining gaps to Limitations, including the scale difference to industrial RMs and the headroom in the head itself.

3.3 Test-Time Scaling: Two Axes

Response-axis scaling: Best-of-N selection. Along the response axis, the RM scores N can-

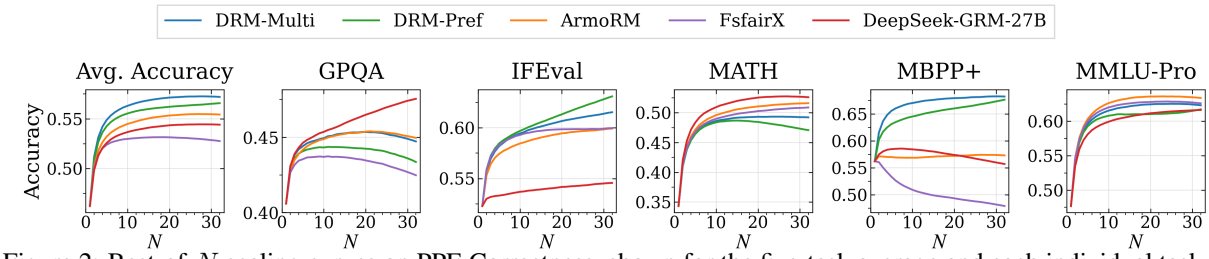

Figure 2: Best-of-N scaling curves on PPE Correctness, shown for the five-task average and each individual task.

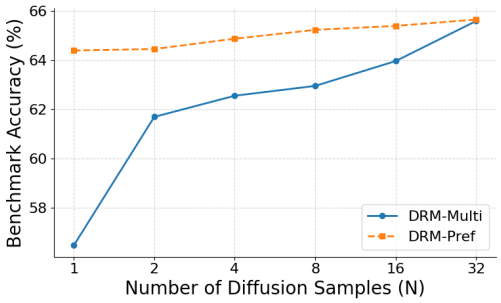

Figure 3: Reward-axis scaling on RewardBench v2.

didate responses and selects the highest, a lever shared by any scalar RM. We evaluate two DRM variants with ArmoRM, FsfairX, and DeepSeek-GRM-27B on PPE Correctness. Figure 2 shows BoN curves on five challenging PPE Correctness tasks together with their average. Averaged over the five tasks, both DRM variants scale best among all evaluated models, and their accuracy improves monotonically as N grows. In contrast, some baselines plateau or even drop at large N, where the highest score may pick the wrong answer, which is a reward-hacking trend that DRM may avoid.

Reward-axis scaling: sampling the reward distribution. The reward axis is unique to DRM: for a fixed (x, y), drawing more samples from the diffusion head tightens the estimate of its reward distribution. As shown in Figure 3, DRM-Multi-8B rises monotonically from 56.5% at a single sample to 65.6% at 32 samples on RewardBench v2, while DRM-Pref-8B improves from 64.4% to 65.7%. The number of diffusion samples thus acts as a test-time knob that trades computation for scoring precision without retraining, a lever no scalar or single-pass parametric RM provides.

3.4 Fine-Grained Analysis of DRM

Balanced preference. In RMB Best-of-N evaluation (Table 5), DRM maintains high and closely matched scores on both Helpfulness and Harmlessness, with an average difference of less than 5 points. In contrast to some reward models that often score high on one dimension but noticeably lower on the other, DRM's balanced performance indicates that its reward representations can encode multimodal preferences simultaneously. Moreover, the two DRM variants show highly consistent performance across both dimensions, further suggesting that this balance reflects the intrinsic multimodal preference modeling capability of the method rather than a coincidental outcome.

Correctness preference judgment. We use JudgeBench (Table 6) and PPE Correctness (Table 7) to analyze whether DRM can capture objective correctness. DRM-Multi-8B achieves an overall score of 58.6 on JudgeBench, outperforming GPT-4o and Claude-3.5-Sonnet. On PPE Correctness, DRM-Multi-8B also reaches 63.8, surpassing multiple open-source reward models and maintaining stable performance on subtasks such as math, MMLU, and MBPP. These results indicate that the reward representations learned by DRM do not merely reflect response style, politeness, or general helpfulness, but can also identify correctness differences in factual, knowledge, mathematical, and code to some extent. This is important for reward models, since real world preferences often involve both subjective quality and objective correctness.

3.5 What a Scalar Head Cannot Represent

In this section, we illustrate our central claim that human preference is multimodal and that a non-parametric head can represent what a scalar or unimodal head cannot, on a single prompt–response pair. We collect 100 scores each from DRM, a discriminative RM (ArmoRM), and a generative RM (DeepSeek-GRM-27B), compared on a common reward dimension and normalized to $[-1, 1]$. The three expose uncertainty differently: DRM's 100 scores are independent draws from the distribution its diffusion head explicitly models; ArmoRM is deterministic and collapses to a single point; DeepSeek-GRM varies only through decoding temperature, i.e. noise in generating a verdict rather than a modeled reward distribution.

As shown in Figure 4, DRM places mass on

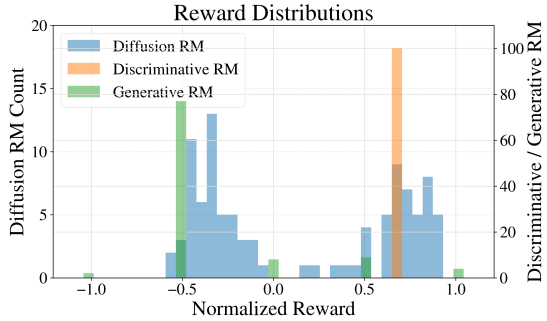

Figure 4: Reward distributions of diffusion, discriminative, and generative reward models.

several well-separated reward regions, recovering a clearly multimodal distribution, while ArmoRM concentrates at a single point and DeepSeek-GRM remains sharply unimodal. This is exactly what the framework predicts: heads that emit a scalar, or that derive randomness only from decoding, commit to one dominant judgment per pair, whereas DRM retains the multiple defensible verdicts a contested example admits. As a single-example illustration this evidence is qualitative, but it shows directly the multimodal reward structure that motivates DRM and that conventional heads cannot express.

4 Related Work

Reinforcement Learning from Human Feedback. RLHF (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022) is the standard recipe for post-training LLMs, spanning PPO (Schulman et al., 2017), GRPO (Shao et al., 2024), reward-ranked fine-tuning (Dong et al., 2023), and direct preference methods such as DPO (Rafailov et al., 2023), all relying on a learned reward to fit human feedback. RLVR-style training with verifiable rewards (Lambert et al., 2024; Guo et al., 2025) is an exception in domains with ground-truth answers, but in general-domain alignment reward quality bounds alignment quality (Gao et al., 2023; Casper et al., 2023). Rather than proposing a new RL algorithm, we ask what form the learned reward should take.

Multimodal Structure of Human Preference. A growing body of work documents that human preference is multimodal instead of a single point. Inter-annotator agreement on RLHF datasets is only 63–73% across Anthropic-HH (Bai et al., 2022), OpenAI summarization (Stiennon et al., 2020), and InstructGPT (Ouyang et al., 2022), with most disagreement reflecting genuine individual preferences rather than noise (Zhang et al., 2024); preference also varies across cultures and demographics (Kirk et al., 2024), and alignment further reduces this diversity (Sorensen et al., 2024). Decomposing reward into interpretable attributes (Wang et al., 2024a,c) surfaces part of this structure, but the distribution over each attribute remains non-degenerate, which a reward model should capture rather than average away.

The Paradigm of Reward Modeling. Reward models follow three paradigms. **Discriminative** RMs regress a scalar over preference pairs with BT loss (Bradley and Terry, 1952; Stiennon et al., 2020; Ouyang et al., 2022), scaled through large preference mixtures (Liu et al., 2025a; Cui et al., 2023; He et al., 2025; Wang et al., 2026) and multi-attribute regression (Wang et al., 2024a,c). **Generative** RMs emit a textual verdict or critique, from prompting frontier LLMs (Zheng et al., 2023; Gu et al., 2024) to reasoning-trained and rubric-based judges (Guo et al., 2026; Chen et al., 2025b; Gunjal et al., 2025; Viswanathan et al., 2026), trading inference cost for interpretability but still producing a scalar score per query. A third line replaces the scalar head with a **parametric distributional** one: DPL (Siththaranjan et al., 2024) and URM (Lou et al., 2024) predict a Gaussian, DPRM (Li et al., 2024) a categorical distribution via optimal transport, and QRM (Dorka, 2024) a fixed quantile grid, each fixing a distribution family in advance and limiting the shapes $p(\mathbf{r} \mid x, y)$ can take. We instead use a Diffusion Transformer (Peebles and Xie, 2023) as the reward head, representing $p(\mathbf{r} \mid x, y)$ within a single architecture supporting both multi-attribute denoising and a distributional BT objective.

5 Conclusion

We presented **DRM**, a reward model whose DiT head models the conditional reward distribution $p(\mathbf{r} \mid x, y)$ without committing to any parametric family. A single architecture handles both multi-attribute and preference data, and at inference yields distributional statistics and a reward-axis scaling lever unavailable to scalar RMs. Across five benchmarks, DRM is competitive with substantially larger RMs under a modest training budget and recovers multimodal reward structure that conventional heads cannot. Future work includes adapting the diffusion head to better absorb pairwise preference signal, scaling to larger curated data, and validating DRM's distributional statistics in downstream RLHF.

Limitations

This work is an initial exploration of a new reward-modeling paradigm, and several limitations remain. First, DRM is trained on only a few hundred thousand open-source samples, far below the scale of industrial RMs trained on tens of millions of curated preferences, and does not yet match the strongest open-source scalar RMs in absolute terms. Second, the DiT reward head itself is far from optimized: DRM-Pref-8B does not yet surpass DRM-Multi-8B despite using more recent high-quality preference data, indicating that the current head underexploits pairwise preference signal and that its architecture and training objective leave clear room for improvement. Third, our study fixes a single 8B encoder and a fixed training-data scale, so how DRM behaves across backbone sizes, model families, and larger data regimes remains unexplored; characterizing these scaling trends is an important direction for establishing the approach.

Moreover, like other reward models, DRM may inherit biases from its training data and could assign high scores to outputs that are stylistically persuasive but factually incorrect or socially harmful. If used as an optimization objective for downstream language models without human oversight or safety constraints, it may amplify such biases or encourage reward hacking behaviors.

References

Anthropic. 2024. Claude 3.5 Sonnet.

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.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and 1 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Shikai Chen, Jin Yuan, Yang Zhang, Zhongchao Shi, Jianping Fan, Xin Geng, and Yong Rui. 2025a. LDL-Reward-Gemma-2-27B-v0.1. Label Distribution Learning for Reward Modeling. Tech report forthcoming.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, and 1 others. 2025b. Rm-r1: Reward modeling as reasoning. *arXiv preprint arXiv:2505.02387*.

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, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, and 1 others. 2023. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

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*.

Nicolai Dorka. 2024. Quantile regression for distributional reward models in rlhf. *arXiv preprint arXiv:2409.10164*.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2025. How to evaluate reward models for rlhf. In *International Conference on Learning Representations*, volume 2025, pages 18128–18163.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. 2025. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Jiaxin Guo, Zewen Chi, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. 2026. Reward reasoning models. *Advances in Neural Information Processing Systems*, 38:150477–150510.

Bingxiang He, Wenbin Zhang, Jiaxi Song, Cheng Qian, Zixuan Fu, Bowen Sun, Ning Ding, Haiwen Hong, Longtao Huang, Hui Xue, and 1 others. 2025. Air: A systematic analysis of annotations, instructions, and response pairs in preference dataset. *arXiv preprint arXiv:2504.03612*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hannah R Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, and 1 others. 2024. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Dexun Li, Cong Zhang, Kuicai Dong, Derrick Goh Xin Deik, Ruiming Tang, and Yong Liu. 2024. Aligning crowd feedback via distributional preference reward modeling. *arXiv preprint arXiv:2402.09764*.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and 1 others. 2025a. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *arXiv preprint arXiv:2507.01352*.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2025b. Rm-bench: Benchmarking reward models of language models with subtlety and style. In *International Conference on Learning Representations*, volume 2025, pages 44323–44355.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025c. Inference-time scaling for generalist reward modeling. *Preprint*, arXiv:2504.02495.

Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. 2024. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *arXiv preprint arXiv:2410.00847*.

Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. Generative reward models. *arXiv preprint arXiv:2410.12832*.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2025. Rewardbench 2: Advancing reward model evaluation. *arXiv preprint arXiv:2506.01937*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. 2024. Offsetbias: Leveraging debiased data for tuning evaluators. *Preprint*, arXiv:2407.06551.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205.

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.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. 2024. Distributional preference learning: Understanding and accounting for hidden context in rlhf. In *International Conference on Learning Representations*, volume 2024, pages 27448–27471.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020a. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020b. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, and 1 others. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

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

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. In *International Conference on Learning Representations*, volume 2025, pages 63277–63303.

Vijay Viswanathan, Yanchao Sun, Xiang Kong, Meng Cao, Graham Neubig, and Sherry Wu. 2026. Checklists are better than reward models for aligning language models. *Advances in Neural Information Processing Systems*, 38:114728–114754.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024a. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024b. Helpsteer2-preference: Complementing ratings with preferences. *arXiv preprint arXiv:2410.01257*.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh N Sreedhar, and Oleksii Kuchaiev. 2024c. Helpsteer 2: Open-source dataset for training top-performing reward models. *Advances in Neural Information Processing Systems*, 37:1474–1501.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, and 1 others. 2024d. 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)*, pages 3371–3384.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. 2026. Helpsteer3-preference: Open human-annotated preference data across diverse tasks and languages. *Advances in Neural Information Processing Systems*, 38.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. 2024. Regularizing hidden states enables learning generalizable reward model for llms. *arXiv preprint arXiv:2406.10216*.

Lifan Yuan, Ganqu Cui, Hanbin Wang, Ning Ding, Xingyao Wang, Jia Deng, Boji Shan, Huimin Chen, Ruobing Xie, Yankai Lin, Zhenghao Liu, Bowen Zhou, Hao Peng, Zhiyuan Liu, and Maosong Sun. 2024. Advancing llm reasoning generalists with preference trees. *Preprint*, arXiv:2404.02078.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Seyed Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2025. Generative verifiers: Reward modeling as next-token prediction. In *International Conference on Learning Representations*, volume 2025, pages 12476–12505.

Michael JQ Zhang, Zhilin Wang, Jena D Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. 2024. Diverging preferences: When do annotators disagree and do models know? *arXiv preprint arXiv:2410.14632*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Enyu Zhou, Guodong Zheng, Binghai Wang, Zhiheng Xi, Shihan Dou, Rong Bao, Wei Shen, Limao Xiong, Jessica Fan, Yurong Mou, and 1 others. 2025. Rmb: Comprehensively benchmarking reward models in llm alignment. In *International Conference on Learning Representations*, volume 2025, pages 26543–26589.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. 2023. Starling-7b: Improving llm helpfulness & harmlessness with rlaif.

A Implementation Details of Diffusion Reward Head

This section details the internal structure of the Diffusion Reward Head summarized in Section 2.1. The head takes as input a noisy reward $\mathbf{r}_t \in \mathbb{R}^K$, a diffusion timestep t, the textual conditioning representation $\mathbf{h} \in \mathbb{R}^{d_{\text{enc}}}$ produced by the frozen backbone, and a reward mask $\mathbf{m} \in \{0,1\}^K$, and outputs a noise prediction $\hat{\boldsymbol{\epsilon}} \in \mathbb{R}^K$.

Input projection. The noisy reward $\mathbf{r}_t$ is first concatenated with the reward mask $\mathbf{m}$ along the feature dimension, then projected into the DiT hidden space through a linear layer:

$$\mathbf{z}_r = W_r[\mathbf{r}_t; \mathbf{m}] + \mathbf{b}_r.$$

A learnable positional embedding $\mathbf{p}$ is added to obtain the input state of the first DiT block:

$$\mathbf{z}_0 = \mathbf{z}_r + \mathbf{p}.$$

Timestep embedding. The diffusion timestep t is first encoded through a standard sinusoidal embedding and then passed through a two-layer MLP that projects it to the DiT hidden dimension:

$$\mathbf{e}_t = \text{MLP}_t(\text{Sinusoidal}(t)).$$

This design lets the model identify the current denoising stage and adjust its behavior accordingly across timesteps.

Textual conditioning. The textual representation $\mathbf{h}$ is mapped to the same hidden dimension through a separate two-layer MLP:

$$\mathbf{e}_h = \text{MLP}_h(\mathbf{h}).$$

The timestep and textual conditions are then fused additively into a unified conditioning vector:

$$\mathbf{c} = \mathbf{e}_t + \mathbf{e}_h.$$

Conditioning via adaLN-Zero. Within each DiT block, the conditioning vector $\mathbf{c}$ is used to generate modulation parameters via adaptive layer normalization (adaLN-Zero) (Peebles and Xie, 2023), which controls both the self-attention and feed-forward sublayers. This injects timestep and textual information through parameter modulation rather than explicit concatenation, preserving the stability of the DiT architecture. We adopt the adaLN-Zero initialization scheme, in which the modulation and output projection layers are initialized close to zero so that the model begins training near an identity mapping and converges more smoothly.

Output head. After the final DiT block, a linear layer maps the hidden representation back to the K-dimensional reward space to produce the noise prediction $\hat{\boldsymbol{\epsilon}}$.

B Benchmark Details

We describe the five evaluation benchmarks below.

- RewardBench v2. RewardBench v2 is a multi-capability evaluation benchmark for reward models, containing 1,865 test samples. Each sample consists of a prompt, a preferred response, and multiple rejected responses. The benchmark mainly adopts a best-of-4 evaluation format to assess whether a reward model can select the best response. It covers 6 task categories, including factuality, precise instruction following, mathematics, safety, focus, and ties, and is used to evaluate the corresponding capability of reward models.
- PPE Preference + PPE Correctness. PPE consists of 2 subsets: PPE Preference and PPE Correctness. PPE Preference contains 16,038 preference pairs. PPE Correctness contains 2,555 prompts, each paired with 32 model responses, and provides correctness labels based on tasks with verifiable answers, including MMLU-Pro, MATH, GPQA, IFEval, and MBPP-Plus. This benchmark supports Best-of-N evaluation and can be used to assess both the model's alignment with real human preferences and its ability to identify objectively correct responses.
- RMB. RMB is a comprehensive benchmark for evaluating reward models, primarily constructed around two alignment objectives: helpfulness and harmlessness. Its samples include both pairwise preference data and Best-of-N data. The significance of RMB lies in the fact that it evaluates not only the pairwise preference judgment ability of reward models, but also their practical capability for candidate selection during inference-time scaling and alignment optimization.
- RM-Bench. RM-Bench contains 1,327 test samples, where each sample consists of one prompt, three chosen responses, and three rejected responses. RM-Bench defines three difficulty levels: easy, normal, and hard. The data cover domains including chat, code, math, safety refusal, and safety response. This benchmark primarily evaluates a reward model's sensitivity to subtle

content differences, as well as whether it can still identify genuinely better responses under interference from response length, formatting, and stylistic variation.

- JudgeBench. JudgeBench is an evaluation benchmark for LLM judges and reward models, containing 350 response pairs generated by GPT-4o and 270 response pairs generated by Claude-3.5-Sonnet. Each sample consists of a question, two candidate responses, and a preference label based on objective correctness. The data cover tasks such as knowledge, reasoning, mathematics, and coding. Unlike evaluations that primarily rely on subjective human preferences, JudgeBench places greater emphasis on factual and logical correctness, and can therefore be used to assess whether a reward model can identify the objectively more correct response in complex response pairs.

C Hyperparameter Tuning

We tune the architecture and optimization settings of the Diffusion Reward Head and then fix the selected configuration across all benchmarks to avoid tuning on the test sets. The search is conducted on RewardBench v2 (Malik et al., 2025), with the search space summarized in Table 3. The tuned dimensions are the DiT hidden size, number of Transformer blocks, number of attention heads, dropout rate, learning rate, batch size, and the diffusion beta schedule; for DRM-Pref-8B we additionally tune the Bradley-Terry loss weight λ_{BT}.

Hyperparameter	Search range
Hidden size	$\{384, 512, 768\}$
Transformer blocks	$\{3, 4, 5\}$
Attention heads	$\{6, 8, 12\}$
Dropout	$\{0.0, 0.1, 0.2\}$
Learning rate	$\{0.5, 1, 2, 4\}\times 10^{-4}$
Batch size	$\{64, 128, 256\}$
Beta schedule	{linear, sqcos}
λ_{BT} (Pref only)	$\{0.1, 0.5, 1.0\}$

Table 3: Hyperparameter search space.

Our search was conducted in two stages:

1. Core architecture parameter search: We explored the impact of DiT hidden size, number of Transformer blocks, number of attention heads, and the diffusion beta schedule on performance.

2. Training parameter fine-tuning: With the core architecture fixed, we searched for all feasible combinations of learning rate, dropout, and batch size to optimize convergence speed and training stability.

Hidden	Depth	Heads	Beta schedule	RewardBench v2
384	3	6	sqcos	**62.88**
384	4	6	sqcos	59.83
384	5	6	sqcos	62.04
512	4	8	sqcos	62.38
512	4	8	linear	61.76
512	4	8	sqcos	61.84
768	4	12	sqcos	54.80

Table 4: Maximum scores on RewardBench v2 for each core architecture parameter combination (Max Reward). The final selected core architecture configuration is highlighted. All results were obtained with inference settings: num_samples=8, guidance_scale=3.5, num_steps=20.

Table 4 presents the maximum RewardBench v2 scores for all core architecture parameter settings across different training parameter combinations. The final selected configuration, 384_3_6_sqcos, is highlighted. The corresponding training parameters are set to lr = 5e-4, Dropout = 0.2, and Batch = 64. This model achieves the best performance among all combinations while maintaining low training and inference cost, which is the final DRM-Multi-8B model.

After determining the core architecture parameters, we fixed the structure and trained DRM-Pref-8B with different values of λ_{BT}, The value λ_{BT} =0.5, which achieved the best performance on the validation set, was ultimately selected for the DRM-Pref-8B model.

D Other Experiments Details

We provide per-task breakdowns on RMB, JudgeBench, and PPE Correctness in Table 5, Table 6, and Table 7. The RMB results use its Best-of-N splits and are not directly comparable to the RMB pairwise scores in Table 2.

The full scoring pipeline consists of a frozen FsfairX-LLaMA3-RM-v0.1 encoder and a RewardDiT denoiser. The FsfairX encoder has 7.50B parameters, while RewardDiT has approximately 12.0M parameters; therefore, the full encoder-plus-denoiser pipeline contains approximately 7.52B parameters. Only the RewardDiT component is trained.

Experiments were run on NVIDIA A800-SXM4-80GB GPUs using PyTorch 2.7.0+cu126, CUDA 12.6, Hugging Face Transformers/Datasets, and Diffusers. RewardDiT checkpoint training required

Table 5: BoN split of RMB results of different reward models.

Model	Helpfulness (BoN)	Harmlessness (BoN)	Avg.
ArmoRM-Llama3-8B-v0.1	63.6	49.7	56.7
Skywork-Reward-Llama-3.1-8B-v0.2	60.5	56.8	58.7
internlm2-7b-reward	62.6	56.3	59.5
DeepSeek-GRM-27B	63.9	58.0	61.0
Eurus-RM-7b	67.9	54.3	61.1
Claude-3.5-Sonnet-20240620	70.5	51.8	61.2
Skywork-Reward-Gemma-2-27B-v0.2	63.1	59.9	61.5
GPT-4o-20240513	63.9	68.2	66.1
DRM-Multi-8B	66.2	61.4	63.8
DRM-Pref-8B	66.5	61.3	63.9

Table 6: JudgeBench results of different reward models.

Model	Knowledge	Reasoning	Math	Code	Avg.
Arena-Hard/GPT-4o-20240513	51.6	52.6	68.4	49.1	54.3
Arena-Hard/Claude-3.5-Sonnet-20240620	52.3	59.6	61.0	48.3	54.6
DRM-Multi-8B	55.8	62.0	63.7	57.9	58.6
DRM-Pref-8B	54.9	57.0	62.2	60.3	57.1

Table 7: PPE correctness results of different reward models.

Model	PPE-MMLU	PPE-MATH	PPE-GPQA	PPE-IFEval	PPE-MBPP	Avg.
Skywork-Reward-Gemma-2-27B	53.9	62.7	52.7	53.8	59.3	56.5
Eurus-RM-7b	63.4	69.3	53.9	59.4	54.0	60.0
Starling-RM-34B	67.7	66.4	57.0	56.1	54.6	60.3
internlm2-7b-reward	66.7	72.6	54.6	64.0	43.9	60.4
Skywork-Reward-Llama-3.1-8B-v0.2	64.3	69.6	56.5	61.5	51.6	60.7
ArmoRM-Llama3-8B-v0.1	66.5	70.7	57.0	58.4	54.2	61.4
DRM-Multi-8B	66.1	69.9	57.5	60.1	65.3	63.8
DRM-Pref-8B	66.3	69.1	56.2	61.5	59.3	62.5

approximately 1.54 GPU-hours in total, including per-epoch validation: 1.11 GPU-hours for the ArmoRM checkpoint and 0.43 GPU-hours for the Tulu3 pair-preference checkpoint. These estimates exclude the one-time cost of generating text embeddings with the frozen FsfairX encoder.

E Responsible NLP and Artifact Use

Artifact Use and Access Conditions. We use publicly available datasets, benchmarks, pretrained models, and baseline models for research, training, evaluation, and comparison purposes. We cite the original creators of these artifacts and use them in accordance with their respective licenses, terms of use, and access conditions. We do not redistribute artifacts whose licenses or access conditions prohibit redistribution.

Data Privacy and Sensitive Content. We do not collect new personal data or attempt to identify individuals. Our experiments use publicly available datasets and benchmarks. Some preference, safety, or harmlessness-oriented datasets may contain sensitive, toxic, or offensive content as part of their intended research and evaluation scope. We use such data only for research and aggregate evaluation.

LLM Usage. During the preparation of this manuscript, we used large language models (LLMs) to assist with language polishing and improving the clarity and readability of the paper. The LLMs were not used to generate research hypotheses, design the methodology, conduct experiments, analyze results, or draw conclusions. All LLM-assisted edits were carefully reviewed and revised by the authors, who take full responsibility for the final content of the manuscript.