# Evaluating Reward Model Generalization via Pairwise Maximum Discrepancy Competitions

**Anonymous ACL submission**

## Abstract

Reward models (RMs) are central to aligning large language models, yet their practical effectiveness hinges on *generalization* to unseen prompts and shifting distributions. Most existing RM evaluations rely on static, pre-annotated preference datasets, which provide limited coverage and often fail to faithfully assess generalization in open-world settings. We introduce Pairwise Maximum Discrepancy Competition (PMDC), a dynamic and annotation-efficient framework for evaluating RM generalization using a large, unlabeled, open-domain prompt pool. PMDC *actively* selects prompt–response pairs that maximize disagreement between two RMs, yielding a compact set of highly contentious test cases. These cases are adjudicated by an oracle, and the resulting outcomes are aggregated via a Bradley–Terry model to produce a global ranking and pairwise win-rate landscape of RMs. We apply PMDC to re-evaluate 10 representative RMs and observe substantial rank reshuffling compared with conventional benchmarks. Qualitative analyses further uncover systematic generalization failures, providing valuable insights for improving reward modeling.

## 1 Introduction

Reward models (RMs) are a cornerstone of modern alignment pipelines, enabling large language models (LLMs) to internalize complex human preferences through reinforcement learning from human feedback (RLHF) (Christiano et al., 2023; Stiennon et al., 2022). By learning to predict human preferences between pairs of model responses, RMs provide scalable training signals that guide LLMs toward desirable behaviors across various domains, including instruction following, reasoning, and safety (Bai et al., 2022; Ouyang et al., 2022; Feng et al., 2025). Their real-world effectiveness, however, is determined by their ability to generalize to unseen prompts and shifting data distributions.

However, prevailing evaluation benchmarks (Lambert et al., 2024) for RMs predominantly rely on static, pre-annotated datasets offering limited coverage and are inadequate for faithfully assessing generalization in open-world settings. These conventional evaluation resources suffer from several critical limitations. First, their restricted coverage of the potential prompt and behavioral space impedes the assessment of model generalization to novel domains or edge-case scenarios. Second, the human annotations underpinning these datasets are typically sourced from specific demographic groups or constrained task contexts, potentially introducing biases that do not accurately reflect broader human judgment. Third, the fixed and publicly accessible nature of these test sets introduces inherent risks of overfitting, both explicit and implicit, where models may be optimized for benchmark performance without achieving meaningful improvements in alignment quality or robustness (Gao et al., 2022; Zhong et al., 2025; Kim et al., 2025).

To address these challenges, we propose the *Pairwise Maximum Discrepancy Competition* (PMDC), a dynamic and cost-efficient framework for evaluating generalization of RMs. Inspired by discrepancy-driven evaluation paradigms in computer vision (Saito et al., 2018), PMDC actively identifies prompt-response pairs that maximize disagreement between two RMs from a large, unlabeled, open-domain prompt pool, producing a focused set of highly contentious test cases. Such contentious instances are adjudicated by an oracle, here, a powerful LLM serving as a scalable proxy for human judgment. The results are aggregated via a Bradley-Terry (BT) model (BRADLEY and TERRY, 1952) to generate a global ranking and detailed win-rate landscape.

Crucially, PMDC shifts the evaluation paradigm from static benchmarking to active and adaptive probing. This approach offers two key advantages. First, it enables *dynamic evaluation* by adaptively sampling test cases from an open-domain prompt pool and utilizing responses generated by a diverse set of LLMs, thereby facilitating the detection of out-of-distribution failures and enhancing generalization assessment. Second, it ensures *annotation efficiency* by submitting only the most discriminative sample pairs to the oracle for judgment, which significantly reduces annotation costs. We apply PMDC to re-evaluate 10 representative RMs, observing substantial rank reshuffling compared to conventional benchmarks and uncovering systematic generalization failures that provide actionable insights for improving reward modeling.

The main contributions of this work are threefold:

- **The PMDC Framework:** A novel paradigm that moves beyond static benchmarks to dynamically evaluate RM generalization through active, discrepancy-driven sampling.
- **An Actively Probed Dataset:** A compact, high-quality evaluation set, generated by identifying points of maximum inter-model disagreement, which facilitates discrimination-rich assessment and can enhance downstream alignment.
- **New Empirical Insights:** A re-evaluation of state-of-the-art RMs that reveals significant ranking inconsistencies with prior benchmarks and provides valuable diagnostics for generalization failures.

## 2 Related Works

### 2.1 Reward Model Benchmarks

Initial efforts to evaluate RMs established the paradigm of measuring preference prediction accuracy on curated, static datasets. Benchmarks like RewardBench provide foundational, closed-set assessments of RM capabilities (Lambert et al., 2024). However, subsequent research has questioned whether accuracy on such narrow datasets reliably correlates with downstream alignment performance or generalization (Wen et al., 2025; LeVine et al., 2024). In response, more comprehensive benchmarks have emerged, including RM-Bench (Liu et al., 2024b) and RewardBench 2 (Malik et al., 2025), which assess RMs on nuanced capabilities like discerning subtlety and resisting stylistic bias. This paradigm has further extended into specialized domains, reflecting the growing application scope of reward modeling. Recent benchmarks now assess multilingual (Gureja et al., 2025), vision-language (Yasunaga et al., 2025; Li et al., 2025), and embodied agent (Men et al., 2025) RMs. Concurrently, the adoption of powerful LLMs as reward models or preference judges has gained significant traction (Zheng et al., 2023; Dong et al., 2024). To formalize and standardize the assessment of these LLM-based evaluators, several dedicated benchmarks have been proposed (Thakur et al., 2025; Murugadoss et al., 2024; Tan et al., 2025; Zhou et al., 2025). These frameworks evaluate critical dimensions such as alignment with human preferences, robustness to varying instruction complexities, and consistency across diverse evaluation scenarios, establishing much-needed rigor in judge-style model assessment.

### 2.2 Maximum Discrepancy Competition

Beyond static benchmarks, it is crucial to actively and efficiently probe for model weaknesses. The Maximum Discrepancy (MAD) competition framework provides a powerful methodology for this task (Ma et al., 2020). Instead of relying on a fixed test set, MAD adaptively samples data points that cause the largest disagreement between two or more competing models. This principle has been successfully applied to expose failures and compare models in diverse domains, including objective image quality (Ma et al., 2016) and semantic segmentation (Yan et al., 2021). More recently, this sample-efficient approach has been adapted for the human evaluation of large language models, demonstrating its effectiveness in identifying the most informative examples to distinguish between high-performing models (Feng et al., 2025). Our work is inspired by this adversarial, comparative approach to develop a more robust and efficient evaluation protocol for reward models.

## 3 Proposed PMDC

This section details the PMDC, our dynamic framework for evaluating RM generalization. As illustrated in Figure 1, our method proceeds in three core stages: 1) constructing a diverse, unlabeled evaluation pool; 2) actively selecting the most informative, high-discrepancy test cases; and

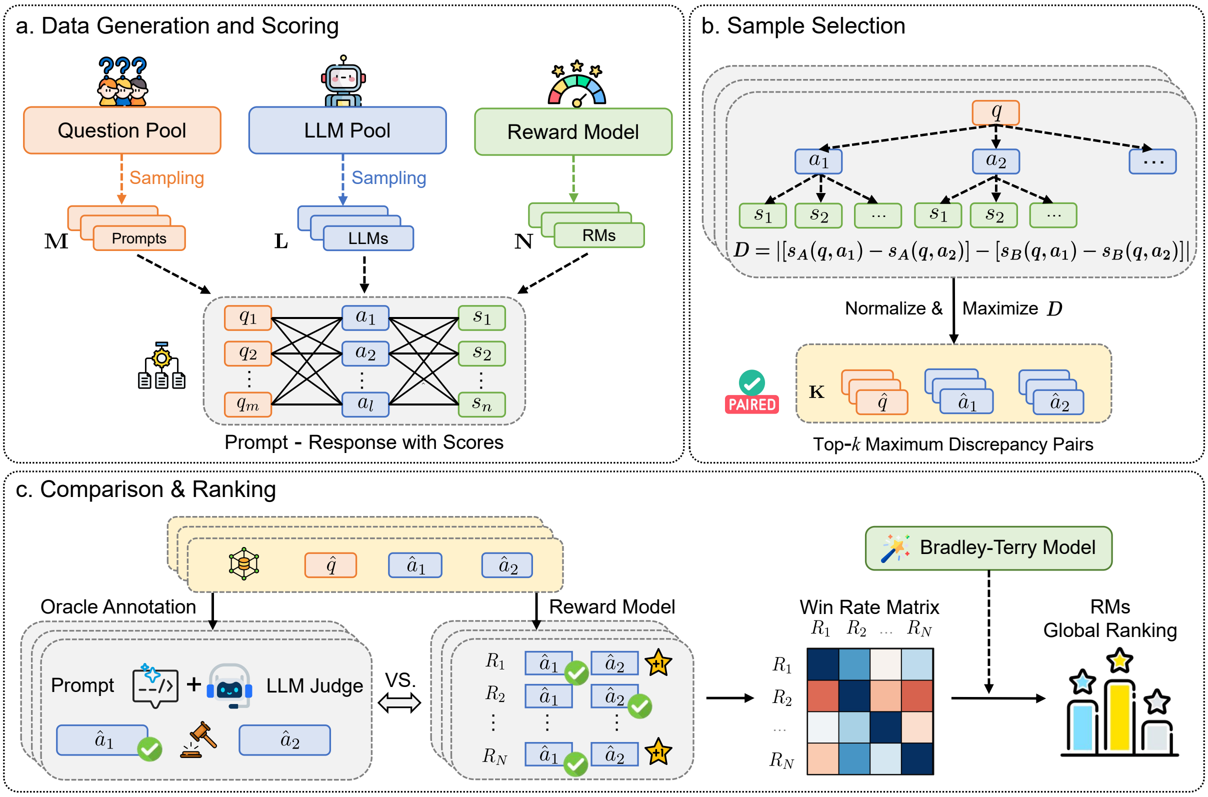


Figure 1: Overview of the proposed PMDC framework. (a) *Data generation and scoring*: Sample prompts and LLMs to build prompt-response pairs, which are then scored by reward models (RMs). (b) *Sample selection*: Based on the Maximum Discrepancy Competition principle, select top-$k$ pairs (with maximum RM preference discrepancy) to form an evaluation subset. (c) *Comparison & ranking*: Annotate the selected QA pairs with an Oracle (i.e., LLM-based Judge) to rank responses, compare Oracle results with RMs to build a win-rate matrix, and convert the pairwise comparisons into RMs' global ranking using the Bradley-Terry model.

3) adjudicating these cases and aggregating the results into a global ranking.

### 3.1 Data Generation and Scoring

We begin by constructing a diverse evaluation corpus $\mathcal{X}$. We sample $M$ prompts from a comprehensive, open-domain pool to ensure broad topical and stylistic coverage. Concurrently, we select $L$ distinct LLMs from a diverse model pool. For each prompt $q_j$, we generate $L$ candidate responses using the selected LLMs, yielding a dataset $\mathcal{X} = \{(q_j, \{a_j^{(m)}\}_{m=1}^{L})\}_{j=1}^{M}$ that captures a wide range of response strategies.

Each prompt-response pair $(q, a)$ is then evaluated by $N$ distinct RMs $\mathcal{R} = \{R_i\}_{i=1}^{N}$. Each RM $R_i$ assigns a real-valued score $s_i(q, a)$ reflecting its assessment of the response's quality. To enable a fair comparison across RMs with potentially different output scales, we apply min-max normalization per model across the entire dataset:

$$s'_i(q, a) = \frac{s_i(q, a) - \min_i}{\max_i - \min_i}, \quad (1)$$

where $\min_i$ and $\max_i$ represent the minimum and maximum scores produced by $R_i$ over the dataset, respectively. These normalized scores are converted into discrete preferences:

$$\text{Pref}(R_i; q, a_1, a_2) = \begin{cases} 1 & \text{if } s'_i(q, a_1) > s'_i(q, a_2), \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

### 3.2 Sample Selection

The objective is to efficiently evaluate and rank $N$ RMs on the dataset $\mathcal{X}$. Conventional evaluation methods heavily rely on static and pre-annotated datasets, which frequently fail to measure the generalization ability of RMs when confronted with unseen prompts in open-world settings. The standard process for evaluating RMs on new datasets consists of three stages. First, a small dataset $\mathcal{S}$ must be pre-selected. Second, predictions are generated by processing $\mathcal{S}$ through the RMs. Third, human evaluation is performed on these outputs to compare relative model performance. The RM that achieves the highest average subjective rating across $\mathcal{S}$ is considered superior. However, this evaluation paradigm is labor-intensive, expensive, and challenging to scale, posing significant practical limitations for efficient assessment of reward models.

Following the MAD principle (Wang and Simoncelli, 2008), we aim to evaluate the RM generalization by adaptively selecting a minimal yet highly informative subset of prompt-response pairs. We begin by considering the *simplest* scenario, where two RMs $R_A$ and $R_B$ are being compared under an oracle budget that permits the judgment of only one prompt and its corresponding response pair $(q, \{a_1, a_2\}) \in \mathcal{X}$. The core challenge thus reduces to: How can we automatically select the most informative sample from a large pool of candidates such that the relative performance between $R_A$ and $R_B$ can be most effectively discerned?

According to the MAD competition methodology, PMDC selects the prompt-response pair $(\hat{q}, \{\hat{a}_1, \hat{a}_2\}) \in \mathcal{X}$ that best differentiates between RMs $R_A$ and $R_B$:

$$(\hat{q}, \{\hat{a}_1, \hat{a}_2\}) = \underset{(q,\{a_1,a_2\})\in\mathcal{X}}{\arg\max} \Big|[s'_A(q,a_1) - s'_A(q,a_2)] - [s'_B(q,a_1) - s'_B(q,a_2)]\Big|. \tag{3}$$

where $s'_A(q,a_1) - s'_A(q,a_2)$ represents the preference score difference assigned by model $R_A$ to the response pair $\{a_1, a_2\}$, with a larger positive value indicating a stronger preference for $a_1$ over $a_2$. The same applies to $s'_B(q,a_1) - s'_B(q,a_2)$ for $R_B$.

Then, we extend this idea to compare $R_A$ and $R_B$ over a small subset $\mathcal{S} \subset \mathcal{X}$ comprising $K$ prompt–response pairs with the highest discrepancy values, as computed by Eq. 3. The $k$-th pair is selected iteratively using:

$$(\hat{q}, \{\hat{a}_1, \hat{a}_2\})^{(k)} = \arg \max_{(q,\{a_1,a_2\})\in\mathcal{X}\setminus\mathcal{S}} \Big|[s'_A(q,a_1) - s'_A(q,a_2)] - [s'_B(q,a_1) - s'_B(q,a_2)]\Big|. \tag{4}$$

where $\mathcal{S} = \{(\hat{q}, \{\hat{a}_1, \hat{a}_2\})\}_{i=1}^{k-1}$ contains the previously chosen $k-1$ pairs. Each newly selected pair is incorporated into $\mathcal{S}$ for subsequent iterations.

### 3.3 Comparison & Ranking

The oracle assessment of the preferences from $R_A$ and $R_B$ for a given pair $(q, \{a_1, a_2\})$ leads two plausible results:

- The oracle's judgment is consistent with that of $R_A$ (or $R_B$). In this case, PMDC successfully identifies the most informative prompt–response pair for discriminating between the two models, thereby enabling a conclusive performance ranking.
- The oracle cannot determine a superior response, which is possible in open-world scenarios. Although the selected prompt-response pair $(\hat{q}, \{\hat{a}_1, \hat{a}_2\})$ may reveal divergent strengths (or weaknesses) of $R_A$ and $R_B$, but contributes less to their relative performance ranking.

Given $N$ RMs, PMDC chooses top-$k$ prompt-response pairs for each of the $\binom{N}{2}$ model pairs, resulting in a final evaluation set $\mathcal{D}$ of size $N(N-1)K/2$. Notably, the size of $\mathcal{D}$ is independent of the size of the input domain $\mathcal{X}$, allowing PMDC to benefit from an expanded $\mathcal{X}$ with broader prompt–response coverage.

For the Oracle assessment, PMDC employs a two-alternative forced choice (2AFC) paradigm. Each prompt–response pair $(q, \{a_1, a_2\}) \in \mathcal{S}$ is presented to the oracle alongside the outputs of two competing RMs, $R_A$ and $R_B$. The oracle is required to select the preferred response. The collected judgments are compiled into an $N \times N$ win-count matrix $W$, where $W_{i,j}$ records the number of votes for $R_i$ and against $R_j$. The symmetrized win rate matrix is computed as:

$$P_{i,j} = \frac{W_{i,j}}{W_{i,j} + W_{j,i} + \varepsilon}, \quad P_{i,i} = 0.5, \tag{5}$$

where $\varepsilon$ is a small smoothing constant, ensuring $P_{i,j} + P_{j,i} \approx 1$ off-diagonal and neutral diagonal.

We employ the BT model to infer the global ranking of $\mathcal{R}$. Specifically, we let $\xi$ be the vector of global ranking scores $[\xi_1, \ldots, \xi_n]$, and define the probability of $R_i$ being preferred over $R_j$ as

$$P_{i,j} = \frac{1}{1 + \exp(\xi_j - \xi_i)}. \tag{6}$$

We estimate the global scores by maximizing regularized log-likelihood with Broyden-Fletcher-Goldfarb-Shanno (BFGS) (Hunter, 2004), and applying L2 penalty ($\lambda = 10^{-6}$) for numerical stability and fixing $\xi_1 = 0$ for identifiability:

$$\log \mathcal{L}(\boldsymbol{\xi}) = \sum_{(i,j)\in\mathcal{C}} \Big[w_{ij} \log P_{i,j} + w_{ji} \log P_{j,i}\Big] - \lambda \sum_{k=2}^{n} \xi_k^2. \tag{7}$$

We summarize the proposed PMDC in Algorithm 1 (See Appendix).

## 4 Experiments

We conduct a comprehensive evaluation of the PMDC framework to assess its capability for dynamically evaluating RM generalization. Section 4.1 details our setup, including datasets, reward models, and evaluation metrics. Section 4.2 presents PMDC's global rankings and compares them with established benchmarks. Section 4.3 analyzes PMDC's sensitivity to key design choices, such as top-$k$ selection, oracle judge, and sampling randomness. Section 4.4 presents qualitative case studies of PMDC-selected comparisons. Finally, Section 4.5 demonstrates how PMDC-identified samples can improve reward models via targeted fine-tuning.

### 4.1 Experimental Setup

**Dataset** We first compile a large, unlabeled prompt pool by aggregating prompts from six LLM benchmarks to ensure broad topical coverage: 1) MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2021), 2) GSM8K (Grade School Math) (Cobbe et al., 2021), 3) HumanEval (Chen et al., 2021), 4) AlpacaEval (Li et al., 2023), 5) TruthfulQA (Lin et al., 2022), 6) HellaSwag (Zellers et al., 2019). To mitigate potential data contamination (e.g., prompts appearing in RM pretraining or preference data), we rewrite prompts via instruction evolution (Zeng et al., 2024), which progressively transforms prompts into more diverse and higher-complexity variants. We further construct a pool of 20 state-of-the-art LLMs spanning multiple model families and providers. For each prompt, we randomly sample several models from this pool to generate candidate responses, enabling flexible composition of prompt–response pairs.

**Reward Models** We evaluate 10 representative RMs, covering different architectures, training paradigms, and parameter scales: 1) Skywork-Reward-Gemma-2-27B (Liu et al., 2024a), 2) QRM-Gemma-2-27B (Dorka, 2024), 3) Reward-Model-Mistral-7B-instruct-unified (Yang et al., 2024), 4) URM-LLaMa-3.1-8B (Lou et al., 2025), 5) ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024), 6) Skywork-Reward-Llama-3.1-8B (Liu et al., 2024a), 7) Skywork-Reward-V2-Qwen3-8B (Liu et al., 2025), 8) Reward-Model-Deberta-v3-large-v2 (OpenAssistant, 2023), 9) Skywork-Reward-V2-Llama-3.2-3B (Liu et al., 2025), 10) Skywork-Reward-V2-Llama-3.1-8B (Liu et al., 2025). Among them, models 1) and 2) are large-scale 27B parameter models that demonstrate superior performance, while models 3)-6) represent mid-scale instruction-tuned variants, and models 7)-10) are recent Skywork-V2 series models with various base architectures.

**Oracle** We employ Claude-Sonnet-4 as our oracle judge via the API endpoint. The oracle uses a systematic prompt design with structured instructions to ensure consistent and reliable judgments. The full prompt can be found in Appendix A.

**Evaluation Metrics** In addition to the global ranking from the BT model, we also report the oracle agreement rate, which measures the proportion of Maximum Discrepancy samples where an RM's preference aligns with the oracle's judgment:

$$\text{Agreement}(R_i) = \frac{1}{|\mathcal{S}_i|}\sum_{\mathbf{s}\in\mathcal{S}_i} \mathbb{I}\big[\text{Pref}(R_i,\mathbf{s}) = \text{Oracle}(\mathbf{s})\big]. \tag{8}$$

where $\mathcal{S}_i$ is the set of maximum discrepancy samples involving RM $R_i$. A higher value indicates that the RM is more reliable when evaluating challenging samples characterized by high inter-model disagreement.

**Implementation Details** For each experiment, we randomly select $1,000$ prompts from the compiled prompt pool. For each selected prompt, we generate 5 candidate responses by randomly sampling five LLMs from the LLM pool. For each reward model pair, we systematically select the top-$k$ QA pairs with the highest reward score discrepancy across all $1,000 \times 5 = 5,000$ QA pairs using our normalized score difference metric (Eq. 3), where $k$ is a configurable hyperparameter (default $k = 10$). With 10 RMs, this yields $\binom{10}{2} \times k = 45 \times 10 = 450$ Maximum Discrepancy samples in total.

### 4.2 Main Results

**Global Ranking Results** Table 1 presents the global ranking of the evaluated models, reporting both their BT ranking score and oracle agreement rates. These results provide a direct assessment of generalization by ranking models based on their performance on actively selected, contentious test cases. As expected, higher BT scores correspond to greater consistency with the

| Model | Rank | BT ranking score | Agreement (%) |
|---|---|---|---|
| Skywork-Reward-Gemma-2-27B | 1 | 2.488 | 91.6 |
| QRM-Gemma-2-27B | 2 | 1.271 | 74.8 |
| Reward-Model-Mistral-7B-instruct-unified | 3 | 0.753 | 65.3 |
| URM-LLaMa-3.1-8B | 4 | 0.065 | 49.9 |
| ArmoRM-Llama3-8B-v0.1 | 5 | 0.000 | 48.3 |
| Skywork-Reward-Llama-3.1-8B | 6 | -0.185 | 44.1 |
| Skywork-Reward-V2-Qwen3-8B | 7 | -0.455 | 37.6 |
| Reward-Model-Deberta-v3-large-v2 | 8 | -0.507 | 36.7 |
| Skywork-Reward-V2-Llama-3.2-3B | 9 | -1.007 | 26.0 |
| Skywork-Reward-V2-Llama-3.1-8B | 10 | -1.059 | 24.9 |

Table 1: Global ranking results. Higher agreement indicates better oracle consistency on contentious Maximum Discrepancy samples.

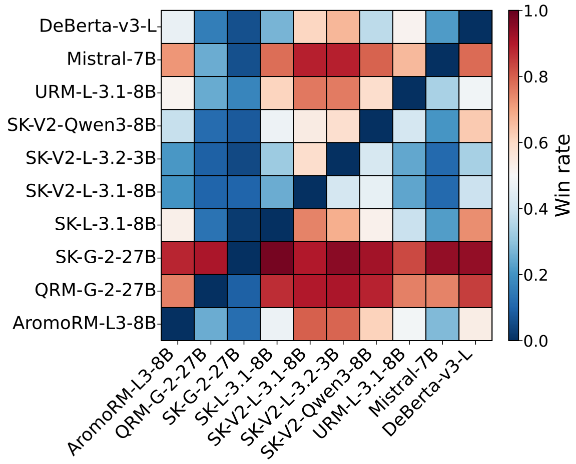


Figure 2: Pairwise win-rate heatmap across RMs on Maximum Discrepancy samples.

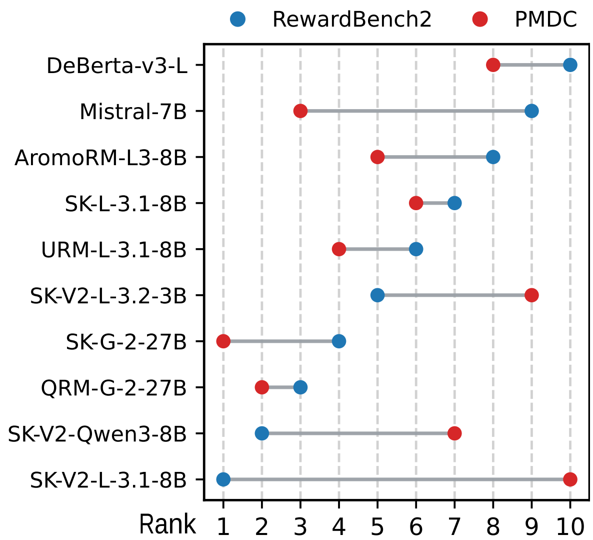


Figure 3: Rank comparison between PMDC and RewardBench2. Horizontal lines connect each reward model's ranking under RewardBench2 (blue) and PMDC (red).

oracle's judgments, confirming the reliability of the estimated preferences. The pairwise win-rate landscape in Figure 2 provides a more detailed view of model performance. The results reveal a clear performance hierarchy with Skywork-Reward-Gemma-2-27B emerging as the top performer, achieving an oracle agreement rate of 91.6% and the highest BT ranking score of 2.488. This indicates consistently strong alignment with human-like judgments across contentious evaluation scenarios. QRM-Gemma-2-27B and Reward-Model-Mistral-7B-instruct-unified follow in second and third place, respectively. The dominance of Gemma-2-27B-based models in the top rankings suggests that scale and architecture significantly influence reward modeling capability. Oracle agreement rates complement the BT ranking by measuring absolute reliability on contentious samples, with the strong positive correlation confirming the robustness of our evaluation framework.

**Comparison Against Established Benchmarks** To assess whether PMDC provides a more faithful evaluation of generalization in open-world settings, we compare its rankings against those from RewardBench2, a challenging held-out evaluation track. This comparison, illustrated in Figure 3, reveals substantial rank reshuffling. While broad trends are consistent for some models, significant discrepancies emerge. A striking example is Skywork-Reward-V2-Llama-3.1-8B, which performs markedly worse under PMDC's evaluation, indicating that static benchmarks may overestimate its robustness when confronted with the

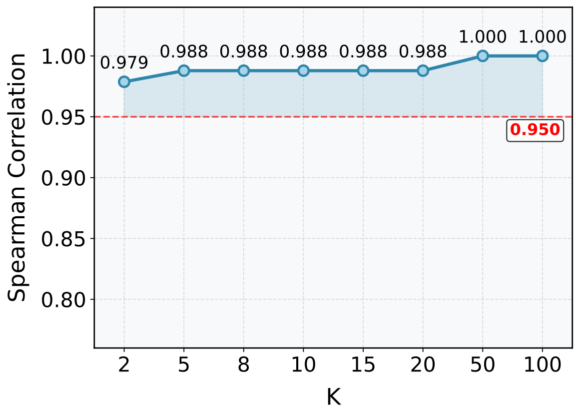


Figure 4: Spearman correlation of PMDC ranks across top-$k$ values. The dashed red line at 0.95 highlights high rank consistency.

novel, contentious scenarios that PMDC actively seeks out (see Appendix F).

This divergence validates PMDC's core methodological premise. Unlike static benchmarks that assess recall on a fixed data distribution, PMDC actively probes for contentious edge cases. By evaluating generalization under distribution shift, it reduces sensitivity to benchmark overfitting and offers a complementary, diagnostics-rich perspective on RM robustness that conventional evaluations often miss.

### 4.3 Model and Result Analysis

**Sensitivity of Top-$k$** To evaluate ranking stability, we vary the number of selected high-discrepancy pairs (top-$k$) from 2 to 100, using the ranking at $k = 100$ as the reference to compute Spearman correlation for each $k$. The results show that even with $k = 5$, the obtained rankings are highly consistent with those using $k = 100$ (see Figure 4). While very low $k$ values increase variance, they can help surface rare edge-case disagreements. In practice, a moderate $k \geq 5$ provides an optimal balance, ensuring both the annotation efficiency and the ranking robustness central to the PMDC framework.

**Annotation Efficiency Analysis** The PMDC framework achieves remarkable annotation reduction compared to exhaustive pairwise evaluation. In our experiments, each of the $1,000$ questions contains 5 candidate responses, requiring $\binom{5}{2} = 10$ pairwise comparisons per question under traditional evaluation, resulting in $1,000 \times 10 = 10,000$ total annotations. In contrast, PMDC with $k = 10$ only requires $\binom{10}{2} \times 10 = 450$ comparisons, a $95.5\%$ reduction in annotation cost. This dramatic efficiency gain demonstrates PMDC's practical utility for large-scale reward model evaluation while preserving ranking fidelity.

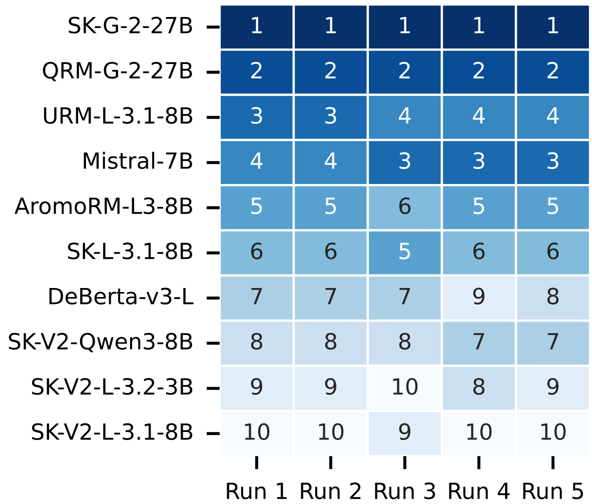


Figure 5: PMDC's rank across five independent runs. The heatmap shows the rank of each RM in each run, with rank values annotated in individual cells.

**Sensitivity of LLM Judge** To assess the robustness of PMDC against potential biases introduced by the choice of oracle, we evaluated the same set of Maximum Discrepancy samples (200 pairs in total) using three distinct LLM judges: Claude-Sonnet-4, Gemini-2.5-Pro, and GLM-4-Plus. In addition, we established a human oracle by asking three NLP researchers to independently evaluate each pair and aggregating their decisions via majority vote. Across LLM judges, the judgements are highly consistent, with agreement rates of $95.0\%$ (Claude vs. Gemini), $96.0\%$ (Claude vs. GLM), and $94.5\%$ (Gemini vs. GLM). Importantly, all LLM judges also exhibit strong agreement with the human oracle, achieving agreement rates of $94.5\%$ (Claude), $92.5\%$ (Gemini), and $93.0\%$ (GLM). These results suggest that PMDC's conclusions are largely insensitive to the specific oracle employed, and that LLM-based judges provide preference signals that closely align with human judgments.

**Result Consistency Analysis** To assess the robustness of PMDC, we conducted five independent evaluation runs using different random seeds for prompt and LLM sampling. As illustrated in Figure 5, the global rankings remain highly stable across all runs, with most models maintaining identical or adjacent positions. In contrast, when using random sampling, i.e., selecting $k = 10$ pairs per model pair randomly from the same pool, the resulting rankings exhibit significantly higher variance across runs, as shown in Figure A1. This instability arises because random samples often fail to capture meaningful points of disagreement between reward models, leading to noisy and inconsistent comparisons. Instead, our PMDC approach produces reliable and reproducible evaluations.

### 4.4 Case Studies

The samples identified by PMDC naturally expose systematic evaluation divergences and contrasting assessment patterns across reward models. We present five representative cases from our empirical evaluation data that reveal how different reward models exhibit distinct preferences and evaluation criteria (see Appendix F). These cases expose critical divergences in how models judge quality across scientific, creative, educational, and technical domains: (1) *Length Bias*: Some models prefer short answers over detailed ones, even when a deeper explanation is needed (Cases 1, 4, 5); (2) *Nuanced Quality Insensitivity*: Some models fail to discern high quality in literary expression, complex reasoning, or specialized technical material (Cases 1, 2, 5); (3) *Context Insensitivity*: Some models use the same standard to judge all responses, even for different kinds of tasks that might need detailed analysis, creative expression, or technical expertise (Cases 1, 2, 3, 5);

These systematic evaluation patterns reveal fundamental differences in how reward models assess content quality, highlighting the importance of careful model selection for domain-specific applications and the need for more nuanced evaluation frameworks that can recognize diverse forms of excellence across diverse domains.

### 4.5 Improving Reward Models via PMDC

Beyond evaluation, PMDC also can improve RMs through targeted fine-tuning. The maximum discrepancy samples identified by PMDC represent precisely those ambiguous or challenging cases where reward models exhibit substantial disagreement, making them ideal candidates for high-leverage fine-tuning. To validate this hypothesis, we train a reward model (baseline) based on ArmoRM-Llama3-8B-v0.1 (Wang et al., 2024) with the same implementation strategy, and fine-tune it using the $4500$ oracle-annotated preference pairs selected by PMDC.

| Dimension | Baseline | Fine-tuned | Gain |
|---|---|---|---|
| Factuality | 0.534 | 0.545 | +1.1% |
| Focus | 0.642 | 0.588 | -5.4% |
| Math | 0.516 | 0.561 | +4.5% |
| Precise IF | 0.366 | 0.381 | +1.5% |
| Safety | 0.780 | 0.764 | -1.6% |
| Ties | 0.570 | 0.641 | +7.1% |
| **Overall** | **0.568** | **0.599** | **+3.1%** |

Table 2: Reward model performance before and after fine-tuning on PMDC-identified samples.

As shown in Table 2, fine-tuning on PMDC-selected samples yields an overall performance gain of $3.1\%$ over the baseline on RewardBench2. The improvement is particularly notable in areas requiring nuanced judgment, such as *Math* and handling of *Ties*. Minor reductions are observed in *Focus* and *Safety*, likely due to the limited size and domain coverage of the fine-tuning set. Nevertheless, the overall improvement demonstrates that discrepancy-driven data selection can effectively boost reward model robustness and alignment fidelity. This result further validates the utility of PMDC not only as an evaluator but also as a data curation engine for reward modeling.

## 5 Conclusion and Future Work

This work introduces the PMDC, a dynamic and efficient framework for evaluating RM generalization, which is often neglected in conventional static benchmarks. By adaptively selecting high-discrepancy response pairs from a diverse and open-domain prompt pool and employing an LLM-based oracle for scalable preference judgment, PMDC enables robust evaluation of RM generalization. Empirical evaluation of 10 RMs not only reveals significant ranking inconsistencies compared to traditional benchmarks but also uncovers nuanced model-specific strengths and weaknesses. These results affirm PMDC's capacity to provide more cost-effective and behaviorally insightful assessments of reward models.

Future research could focus on enhancing oracle reliability through multi-judge ensembles and extending PMDC to multi-dimensional evaluation frameworks to better capture capability trade-offs. Scaling the framework to larger model sets and prompt pools would further enhance its robustness and applicability.

## Limitations

While PMDC offers a cost-effective framework for evaluating RM generalization, several limitations remain. First, unlike conventional benchmarks that evaluate $N$ models independently with $O(N)$ complexity, PMDC relies on pairwise comparisons, resulting in $O(N^2)$ model pairs and additional scoring overhead, which may limit scalability for very large RM collections. Second, PMDC depends on an LLM-based oracle whose inherent biases (e.g., stylistic preferences) may affect evaluation fidelity, though our experiments show high inter-judge agreement. Third, despite efforts to diversify the prompt pool, coverage of highly specialized or rapidly evolving domains remains limited.

## References


Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

RALPH ALLAN BRADLEY and MILTON E. TERRY. 1952. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3-4):324–345.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. *Preprint*, arXiv:1706.03741.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can llm be a personalized judge? *Preprint*, arXiv:2406.11657.

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

Kehua Feng, Keyan Ding, Hongzhi Tan, Kede Ma, Zhihua Wang, Shuangquan Guo, Yuzhou Cheng, Ge Sun, Guozhou Zheng, Qiang Zhang, and Huajun Chen. 2025. Sample-efficient human evaluation of large language models via maximum discrepancy competition. *Preprint*, arXiv:2404.08008.

Leo Gao, John Schulman, and Jacob Hilton. 2022. Scaling laws for reward model overoptimization. *Preprint*, arXiv:2210.10760.

Srishti Gureja, Lester James V. Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2025. M-rewardbench: Evaluating reward models in multilingual settings. *Preprint*, arXiv:2410.15522.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

David R. Hunter. 2004. Mm algorithms for generalized bradley-terry models. *The Annals of Statistics*, 32(1):384–406.

Sunghwan Kim, Dongjin Kang, Taeyoon Kwon, Hyungjoo Chae, Dongha Lee, and Jinyoung Yeo. 2025. Rethinking reward model evaluation through the lens of reward overoptimization. *Preprint*, arXiv:2505.12763.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Rewardbench: Evaluating reward models for language modeling. *Preprint*, arXiv:2403.13787.

Will LeVine, Benjamin Pikus, Anthony Chen, and Sean Hendryx. 2024. A baseline analysis of reward models' ability to accurately analyze foundation models under distribution shift. *Preprint*, arXiv:2311.14743.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. 2025. Vl-rewardbench: A challenging benchmark for vision-language generative reward models. *Preprint*, arXiv:2411.17451.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024a. Skywork-reward: Bag of tricks for reward modeling in llms. *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, Yang Liu, and Yahui Zhou. 2025. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *Preprint*, arXiv:2507.01352.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2024b. Rm-bench: Benchmarking reward models of language models with subtlety and style. *Preprint*, arXiv:2410.16184.

Xingzhou Lou, Dong Yan, Wei Shen, Yuzi Yan, Jian Xie, and Junge Zhang. 2025. Uncertainty-aware reward model: Teaching reward models to know what is unknown. *Preprint*, arXiv:2410.00847.

Kede Ma, Zhengfang Duanmu, Zhou Wang, Qingbo Wu, Wentao Liu, Hongwei Yong, Hongliang Li, and Lei Zhang. 2020. Group maximum differentiation competition: Model comparison with few samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):851–864.

Kede Ma, Qingbo Wu, Zhou Wang, Zhengfang Duanmu, Hongwei Yong, Hongliang Li, and Lei Zhang. 2016. Group mad competition? a new methodology to compare objective image quality models. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1664–1673.

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

Tianyi Men, Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2025. Agent-rewardbench: Towards a unified benchmark for reward modeling across perception, planning, and safety in real-world multimodal agents. *Preprint*, arXiv:2506.21252.

Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2024. Evaluating the evaluator: Measuring llms' adherence to task evaluation instructions. *Preprint*, arXiv:2408.08781.

OpenAssistant. 2023. reward-model-deberta-v3-large-v2.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. *Preprint*, arXiv:1712.02560.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. *Preprint*, arXiv:2410.12784.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *Preprint*, arXiv:2406.12624.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *Preprint*, arXiv:2406.12845.

Zhou Wang and Eero P. Simoncelli. 2008. Maximum differentiation (mad) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):8–8.

Xueru Wen, Jie Lou, Yaojie Lu, Hongyu Lin, Xing Yu, Xinyu Lu, Ben He, Xianpei Han, Debing Zhang, and Le Sun. 2025. Rethinking reward model evaluation: Are we barking up the wrong tree? *Preprint*, arXiv:2410.05584.

Jiebin Yan, Yu Zhong, Yuming Fang, Zhangyang Wang, and Kede Ma. 2021. Exposing semantic segmentation failures via maximum discrepancy competition. *Preprint*, arXiv:2103.00259.

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

Michihiro Yasunaga, Luke Zettlemoyer, and Marjan Ghazvininejad. 2025. Multimodal rewardbench: Holistic evaluation of reward models for vision language models. *Preprint*, arXiv:2502.14191.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Weihao Zeng, Can Xu, Yingxiu Zhao, Jian-Guang Lou, and Weizhu Chen. 2024. Automatic instruction evolving for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6998–7018.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Jialun Zhong, Wei Shen, Yanzeng Li, Songyang Gao, Hua Lu, Yicheng Chen, Yang Zhang, Wei Zhou, Jinjie Gu, and Lei Zou. 2025. A comprehensive survey of reward models: Taxonomy, applications, challenges, and future. *Preprint*, arXiv:2504.12328.

Yilun Zhou, Austin Xu, Peifeng Wang, Caiming Xiong, and Shafiq Joty. 2025. Evaluating judges as evaluators: The jetts benchmark of llm-as-judges as test-time scaling evaluators. *Preprint*, arXiv:2504.15253.

# Appendix

## A Oracle Prompt

**Prompt for Oracle Judge Evaluation**

**System Message:** You are a professional text quality assessment expert. Please carefully compare the quality of two answers, focusing on: 1) Accuracy – whether the information is correct; 2) Relevance – whether it addresses the question; 3) Clarity – whether the expression is clear and understandable; 4) Conciseness – whether it is concise and avoids redundancy; 5) Depth – whether it has insights; 6) Logic – whether it is well-organized; 7) Practicality – whether it is helpful to the questioner. Find the best balance between information content and readability. Only return the result in JSON format, without any explanation.
**User Prompt:** Please judge which of the following two responses is better. Only return the result in JSON format without any explanation.
Question: {question}
Response 1: {response1}
Response 2: {response2}
Please answer strictly in the following JSON format: {"preference": 1} or {"preference": 2}
Where 1 means Response 1 is better, and 2 means Response 2 is better.

## B Random Sampling Results

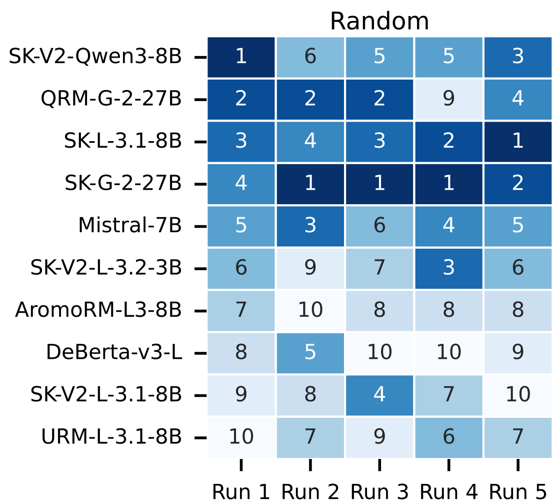


Figure A1: Random sampling's rank across 5 independent runs. The heatmap shows the rank of each RM in each run, with rank values annotated in individual cells.

## C Use of Large Language Models

Large language models were used as assistive tools during the preparation of this manuscript. Specifically, they helped refine language, improve phrasing, and enhance overall readability. All LLM-generated suggestions were carefully reviewed, verified, and edited by the authors.

## D Implementation of PMDC

**Algorithm 1** Pairwise Maximum Discrepancy Competition (PMDC)

**Input:** A Prompt pool, An LLM pool, Reward models $\mathcal{R} = \{R_1, \ldots, R_n\}$, Oracle $\mathcal{O}$, Top-$k$ parameter $k$.
**Initialize:** Pairwise win counts $W_{ij} = 0$ for all $i, j \in \{1, \ldots, n\}$.
Sample a batch of prompts $\mathcal{Q}$ from the Prompt pool
Sample a batch of response generators $\mathcal{G}$ from the LLM pool
**for** each prompt $q$ in $\mathcal{Q}$ **do**
  Generate response set $\{a_1, \ldots, a_m\}$ using generators $\mathcal{G}$.
  Compute all reward scores $s_i(q, a_j)$ for all $R_i \in \mathcal{R}$ and all $a_j$.
  Normalize scores: $s_i'(q, a_j) = \text{Min-Max}(s_i(q, a_j))$ for each model $R_i$.
  **Find Maximum Discrepancy samples:**
  Initialize discrepancy set $\text{MD}_{samples} = \emptyset$.
  **for** each model pair $(R_A, R_B) \in \binom{\mathcal{R}}{2}$ **do**
    Compute discrepancies $D_{A,B}(q, a_i, a_j)$ for all response pairs $(a_i, a_j)$ using Eq. 3.
    $\text{MD}_{\text{pair}} \leftarrow \text{top-}k\{(D_{A,B}, q, a_i, a_j, R_A, R_B) : \forall i < j\}$.
    $\text{MD}_{samples} \leftarrow \text{MD}_{samples} \cup \text{MD}_{\text{pair}}$.
  **end for**
  Total samples: $|\text{MD}_{samples}| = \binom{n}{2} \times k$.
  **Oracle Adjudication:**
  **for** each sample $(D, q, a_i, a_j, R_A, R_B)$ in $\text{MD}_{samples}$ **do**
    Get oracle preference $P_{\mathcal{O}} = \mathcal{O}(q, a_i, a_j)$.
    Get model preferences $P_A = (s_A'(a_i) > s_A'(a_j))$ and $P_B = (s_B'(a_i) > s_B'(a_j))$.
    **if** $P_A = P_{\mathcal{O}}$ and $P_B \neq P_{\mathcal{O}}$ **then**
      $W_{AB} \leftarrow W_{AB} + 1$.
    **else if** $P_B = P_{\mathcal{O}}$ and $P_A \neq P_{\mathcal{O}}$ **then**
      $W_{BA} \leftarrow W_{BA} + 1$.
    **end if**
  **end for**
**end for**
**Global Ranking:**
Compute win-rate matrix $P$ from $W$ using Eq. 5.
Estimate BT ranking score $\boldsymbol{\xi}$ by maximizing Eq. 7 using $W$.
**Output:** Global ranking based on $\boldsymbol{\xi}$.

## E Ethics Statement

This work focuses on evaluating reward models. It does not involve the collection or analysis of personally identifiable, sensitive, or harmful data. All datasets used are publicly available benchmarks or synthetically generated using licensed models, with no private user content included.

## F Cases

Below are five representative cases from our empirical evaluation data that reveal how different reward models exhibit distinct preferences and evaluation criteria.

**Prompt:**
What will happen if one lemming jumps off a cliff?

**Response A (Skywork Preference - Shorter):**
If one lemming jumps off a cliff, the outcome would depend on the specific circumstances, but here's what typically happens:
**1. In Reality**: Contrary to the popular myth, lemmings do not commit mass suicide by jumping off cliffs. This misconception was popularized by the 1958 Disney documentary *White Wilderness*, which staged the scene for dramatic effect.
**Migration & Accidents**: Lemmings may occasionally fall off cliffs or drown while migrating in large groups, especially when crossing unfamiliar terrain or bodies of water. However, this is accidental, not intentional.
**2. If a Single Lemming Falls**: A lemming might survive a short drop if the cliff isn't too high and it lands safely (e.g., in snow or vegetation). However, a significant fall could injure or kill it.
**No Group Effect**: Unlike the myth suggests, one lemming jumping would not trigger others to follow. Lemmings are not blindly obedient creatures.
**3. Ecological Impact**: A single lemming's death would have minimal impact on the population, as they reproduce quickly. However, predators (like owls or foxes) might scavenge the carcass.
**Key Takeaway**: The idea of lemmings mindlessly plunging to their deaths is a myth. In reality, they're simply small rodents that sometimes face hazards during migration—just like other animals.

**Response B (Oracle Preference - Comprehensive):**
This is a fantastic question because it touches on one of the most persistent myths in the animal kingdom. The answer has two parts: the literal, physical reality and the debunking of the famous myth.
**The Literal, Physical Answer**: If one lemming jumps off a cliff, it will fall. What happens next depends entirely on physics: If it's a short drop onto soft ground or into deep water, the lemming might survive. They are decent swimmers. If it's a high drop onto rocks or hard ground, the lemming will almost certainly be injured or killed by the impact.
Crucially, **nothing else will happen. The other lemmings will not mindlessly follow it over the edge**.
**The Myth vs. The Reality**: The popular image of lemmings lining up to commit mass suicide by leaping off cliffs is completely false. This idea was created and popularized by a very specific source.
**Where Did the Myth Come From?**: The myth was cemented in the public consciousness by the 1958 Disney nature film, "White Wilderness." The filmmakers brought lemmings to a cliff, used a spinning turntable to herd and force the frantic animals over the edge, and used clever editing to make it look like a willful, suicidal stampede. The scene was entirely staged.
**What Really Happens in Nature?**: Lemmings have dramatic population cycles. Every 3-4 years, their numbers explode, leading to mass migration to find new food and territory. During these migrations, accidental deaths occur when they try to cross obstacles like rivers and cliffs. These are accidental deaths, not intentional suicide.

**Oracle Judgment**: Response B (comprehensive myth-busting with detailed scientific explanation)
**Model Discrepancy**: Skywork-V2-Llama-3.1-8B (A: 0.907 vs B: 0.319) strongly favored the shorter response, while ArmoRM (A: 0.596 vs. B: 0.745) and Oracle preferred the detailed scientific analysis.

**Response A's Limitations**: While factually correct, it provides only surface-level debunking without explaining the myth's origins or providing comprehensive scientific context. The explanation lacks depth in describing the actual biological behaviors of lemmings.
**Response B's Advantages**: Offers a structured, educational approach with clear separation between physical reality and myth debunking. Provides detailed historical context about the Disney film's staged footage, explains the actual population cycle mechanisms, and distinguishes between accidental deaths and intentional behavior. The response demonstrates superior scientific rigor and educational value.

**Case 1**: Skywork-V2-Llama-3.1-8B vs. ArmoRM-Llama3-8B-v0.1 comparison on scientific myth-busting. The former exhibits bias favoring brevity over comprehensive scientific analysis.

**Prompt:**
Write a funny, interesting, inspiring poem for Women's Day.

**Response A (Oracle Preference - Modern & Bold):**
**"Ode to the Glorious Glitch in the System"**
**Oh, Woman! You magnificent force of nature**,
With your multitasking, your grace, and your *attitude*.
You've changed the world with a laptop, a spatula, and a laser,
While raising kids, running nations, and inventing the *attitude*.
**You were told to sit, so you stood.**
**Told to whisper, so you *roared*.**
Told to stay in the kitchen—
So you turned it into a startup *and* baked a cake *and* ignored.
**You've coded rockets, cured diseases**,
Led revolutions in heels (and sometimes barefoot).
You've balanced budgets, babies, and boardroom decisions,
All while being told, "Smile more, don't be so *ruthless*—wait, no, be *shrewd*?"
**You are not "just" a mother, a wife, or a "lady."**
**You are *also* a storm, a sonnet, a spark, and a scheme.**
You are Beyoncé on stage, Malala in class,
And Aunt Linda who *still* won the neighborhood pickleball team.
So here's to the women who *do*, who *dare*, who *demand*,
**Keep being weird, wise, wild, and wonderful.**
The world doesn't stand a chance.
*(And honestly? It never did.)*

**Response B (V2 Preference - Traditional & Rhyming):**
Here's to the woman, a creature of myth,
Whose handbag contains what the universe smiths.
A charger, a lip balm, a singular sock,
A half-eaten biscuit as hard as a rock,
A receipt from two-thousand-and-three,
And the keys to a car that she sold. Who is she?
She's a master of multitasking arts,
Juggling deadlines and feelings and hearts.
Her brain is a browser with sixty-four tabs,
(Including the one tracking neighborhood cats).
She carries the echo of those who once fought,
The suffragette's spirit that couldn't be bought.
She can be a soft dawn or a hurricane's might,
A comforting whisper or blaze burning bright.
She builds up an empire and builds up a friend,
On her wisdom and strength, you can always depend.
So here's to her laughter, her power, her soul,
The one who makes broken things perfectly whole.
Happy Women's Day! Now go find your crown.
(It's probably under a cushion somewhere in the lounge.)

**Oracle Judgment**: Response A (modern, bold voice with contemporary feminist themes)
**Model Discrepancy**: Skywork-Reward-Llama-3.1-8B (A: 0.288 vs. B: 0.096) favored Response A, while Skywork-Reward-V2-Llama-3.1-8B (A: 0.275 vs. B: 0.703) favored Response B, and Oracle preferred the bold creative approach.

**Response A's Advantages**: Feels more inspiring and empowering while maintaining humor. It takes more creative risks with its structure and language, creating a piece that's both entertaining and genuinely moving. The range of references (from Beyoncé to Malala to "Aunt Linda") makes it more inclusive and celebratory of all types of women. Demonstrates bold creativity and emotional impact.
**Response B's Limitations**: Beautifully crafted but plays it safer. While charming with traditional rhyme schemes and consistent meter, it doesn't reach the same emotional heights or creative boldness as Response A. The divergent model preferences reveal fundamentally different aesthetic philosophies in creative content evaluation.

**Case 2**: Skywork-Reward-Llama-3.1-8B vs. Skywork-Reward-V2-Llama-3.1-8B comparison on creative poetry evaluation. The latter exhibits bias in aesthetic preferences affecting creative content evaluation.

**Prompt:**
Context: [header] How to stop an aggressive cat [title] Make a loud noise to distract the cat. [step] A noise can distract the cat and help redirect it away from its aggressive behavior. You could try clapping your hands, banging two objects, or ringing a bell.

Which ending makes the most sense?
A. [substeps] Making loud sounds to try to distract the cat can also distract it from other, more encouraging behavior. If the cat becomes aggressive, it may become defensive.
B. Direct the cat towards the noise to stop it and help it calm down. [substeps] Common loud noises to attract attention include a roaring vehicle or engine.
C. You could also use your voice, which might get the cat to come near you so that it more easily hears you. [substeps] Give the cat a low-pitched " ha-ha " sound.
D. [substeps] Cats can also be fearful of noises, so only use a loud noise if the cat is acting aggressively and you're afraid it could hurt someone, including another animal. If you use a loud noise to distract the cat, make sure it has a clear escape route in case it wants to flee.

**Response A (QRM-Gemma Preference - Comprehensive Analysis):**
The most logical and complete ending is **D**. Here's why:
**- D provides important context** about when and how to use loud noises to stop an aggressive cat. It warns that cats can be fearful of loud sounds and emphasizes that this method should only be used when necessary (if the cat is a danger to others). It also advises ensuring the cat has an escape route, which is crucial for preventing further stress or aggression.
**- A is contradictory** — it suggests loud noises might distract from "encouraging behavior" but doesn't clarify what that means, and it implies the cat may become defensive, which isn't helpful advice for stopping aggression.
**- B is illogical** — directing a cat *toward* a loud noise would likely increase stress, not calm it down. The example of a "roaring vehicle or engine" is also impractical and overly extreme.
**- C shifts focus to using the owner's voice**, which doesn't align with the original step (using loud noises to distract). The suggested "ha-ha" sound is unclear and not a standard cat behavior modification technique.
**Thus, D is the best choice** because it offers practical, safety-conscious advice while staying relevant to the original step.

**Response B (Skywork Preference - Similar but with Style Elements):**
The most logical and contextually appropriate ending is:
**D. [substeps] Cats can also be fearful of noises...**
**Reasoning:**
- The original step recommends using a loud noise (like clapping or ringing a bell) to **distract** an aggressive cat.
- Option D provides a thoughtful, responsible follow-up: it acknowledges the potential fear response in cats, advises caution, and emphasizes safety for both people and the cat.
- It aligns well with humane behavior management by suggesting the noise be used only when necessary and ensuring the cat can retreat, reducing further stress or escalation.
The other options are less appropriate:
- **A** is confusing and contradictory — distracting from "encouraging behavior" doesn't fit the context.
- **B** suggests directing the cat *toward* the noise, which is counterintuitive and potentially dangerous; also, a roaring engine is impractical and overly intense.
- **C** introduces a strange "ha-ha" sound that doesn't align with standard cat behavior advice and shifts focus oddly to attracting the cat.
Thus, **D** is the best choice.

**Oracle Judgment**: Response B (clearer structure, better formatting, concise analysis)
**Model Discrepancy**: QRM-Gemma-2-27B (A: 0.817 vs. B: 0.883) favored Response A, while Skywork-Reward-V2-Llama-3.1-8B (A: 0.760 vs. B: 0.227) strongly favored Response A.

**Response A's Limitations**: While providing accurate systematic analysis and good breakdown of each option, it lacks the organizational clarity and precise language of Response B. The presentation is more dense and harder to scan, despite containing solid analytical content. The divergent model preferences reveal Skywork's systematic bias against well-formatted educational content, failing to recognize how superior presentation enhances learning effectiveness.
**Response B's Advantages**: Demonstrates superior organization with clear formatting, bold headers, and structured presentation that enhances readability. Uses more precise language ("contextually appropriate", "thoughtful, responsible follow-up") and explicitly mentions "humane behavior management" showing deeper understanding. The checkmark emoji and clean separation of reasoning sections make it easier to follow while maintaining analytical thoroughness.

**Case 3**: QRM-Gemma-2-27B vs. Skywork-Reward-V2-Llama-3.1-8B comparison on educational guidance assessment. The latter exhibits bias against comprehensive analytical responses.

**Prompt:**
Describe a scenario where artificial intelligence could be used to improve the quality and efficiency of healthcare delivery.

**Response A (Gemma Preference - Focused Narrative):**
Certainly! Here's a detailed scenario illustrating how artificial intelligence (AI) can enhance healthcare quality and efficiency:
**Scenario: AI-Powered Personalized Patient Management in a Hospital Setting**
**Context:** A large urban hospital aims to improve patient outcomes and optimize resource allocation in its cardiology department, which handles a high volume of patients with varying degrees of heart disease severity.
**How AI is Used:**
**1. Early Risk Detection and Diagnosis:** Upon patient admission, AI algorithms analyze electronic health records (EHR), including medical history, lab results, imaging scans (like echocardiograms), and even wearable device data. The AI model identifies patients at high risk of adverse cardiac events by spotting subtle patterns that might be missed by clinicians.
**2. Personalized Treatment Planning:** Based on the risk assessment, the system suggests personalized treatment plans. For instance, it recommends medication adjustments, lifestyle interventions, or scheduling for specific diagnostic tests tailored to the patient's unique profile.
**3. Real-Time Monitoring and Alerts:** Wearable devices monitor patients' vital signs continuously. AI detects anomalies such as arrhythmias or sudden changes in blood pressure and immediately alerts healthcare providers.
**Outcomes:** Improved Patient Outcomes, Increased Efficiency, Cost Savings, Enhanced Patient Experience.

**Response B (Oracle & V2-Qwen Preference - Comprehensive Structure):**
Artificial intelligence (AI) has the potential to significantly enhance healthcare delivery by improving quality, efficiency, and accessibility. Here's a detailed scenario illustrating how AI could be applied:
**Scenario: AI-Powered Early Disease Detection and Personalized Treatment in Primary Care**
**1. Patient Data Integration and Analysis** - A patient visits their primary care physician with vague symptoms like fatigue, weight loss, and occasional dizziness. The clinic uses an **AI-driven electronic health record (EHR) system** that aggregates data from past visits, lab results, wearable devices, and genetic information.
**2. Early Diagnosis and Risk Prediction** - The AI detects subtle irregularities and flags a **high risk for Type 2 diabetes and cardiovascular disease**, even though traditional diagnostic thresholds haven't yet been crossed.
**3. Personalized Treatment Planning** - The AI suggests **personalized lifestyle modifications** and **precision medicine interventions** based on the patient's genetic profile.
**4. Continuous Monitoring and Adaptive Care** - Real-time data feeds enable **automatic scheduling** and **prediction of potential complications**.
**5. Streamlining Administrative Tasks** - AI automates **appointment scheduling, insurance pre-authorizations, and billing**.
**Benefits:** Early Detection, Reduced Costs, Personalized Care, Efficiency, Scalability.
**Challenges:** Data privacy, algorithmic bias, human oversight requirements.

**Oracle Judgment**: Response B (comprehensive structure, broader scope, practical considerations)
**Model Discrepancy**: Skywork-Reward-Gemma-2-27B (A: 0.622 vs. B: 0.912) strongly favored Response B, while Skywork-Reward-V2-Qwen3-8B (A: 0.923 vs. B: 0.383) strongly favored Response A with massive preference score discrepancy.

**Response A's Limitations**: While providing coherent cardiology-focused content, it demonstrates significant shortcomings in scope and depth. Offers generic references to "subtle patterns" and "anomalies" without concrete specificity. Lacks consideration of implementation challenges, administrative complexities, or ethical considerations. The narrow departmental focus misses broader healthcare system realities and presents an overly idealistic view without acknowledging practical barriers.
**Response B's Advantages**: Demonstrates superior content depth with comprehensive 5-stage coverage from data integration to administrative efficiency. Provides concrete specificity (actual symptoms, biomarkers, technologies) and excellent hierarchical organization with clear headers. Shows sophisticated understanding of healthcare complexities including privacy concerns, algorithmic bias, and human oversight requirements. Offers balanced perspective addressing both benefits and implementation challenges, making it more valuable for practical healthcare AI deployment.

**Case 4**: Skywork-Reward-Gemma-2-27B vs. Skywork-Reward-V2-Qwen3-8B comparison on healthcare AI applications. The latter exhibits bias against comprehensive technical content and practical implementation considerations.

| Prompt: |
|---|
| What are the wavelengths of light emitted by metals depending on their temperature? |

| **Response A (Skywork Preference - Traditional Academic):** | **Response B (Oracle & ArmoRM Preference - Interactive Educational):** |
|---|---|
| The wavelengths (or spectrum) of light emitted by metals depending on their temperature are primarily described by **blackbody radiation** principles. When a metal is heated, it emits electromagnetic radiation across a range of wavelengths, and the distribution of this emitted radiation depends mainly on its temperature.<br>**Key Concepts:**<br>**1. Blackbody Radiation** - Metals, when heated enough, approximate a *blackbody radiator*—an idealized physical body that absorbs all incident electromagnetic radiation and re-radiates energy characteristic of its temperature.<br>**2. Planck's Law** - The spectral radiance B($\lambda$,T) of a blackbody at temperature T (in kelvins) as a function of wavelength $\lambda$ is given by Planck's radiation law: [Complex mathematical formula with h, c, k_B constants]<br>**3. Wien's Displacement Law** - The wavelength $\lambda$_max at which the radiation is most intense (peak wavelength) is inversely proportional to the temperature: $\lambda$_max = b/T where b $\approx 2.898 \times 10^{-3}$ m·K is Wien's displacement constant.<br>**Typical Wavelengths at Different Temperatures:** At ~500 K (227°C): Peak emission in infrared (around 5.8 $\mu$m), invisible to human eye. Around 1000 K (~727°C): Peak shifts to near-infrared (~2.9 $\mu$m), some thermal glow begins. At 1500 K (~1227°C): Peak near 1.9 $\mu$m, metal starts to glow dull red. At 2000 K (~1727°C): Peak around 1.45 $\mu$m, glowing brighter red-orange. At 3000 K (~2727°C): Peak shifts near 1 $\mu$m, metal glows bright orange-white. At 5000 K (~4727°C): Peak near 0.58 $\mu$m (yellow-green), very bright white light.<br>**Practical Notes:** Real metals do not emit as perfect blackbodies. Their emissivity (how efficiently they emit radiation) varies with wavelength and material. Thermal radiation spectrum is continuous, unlike atomic emission lines. | Of course! This is an excellent question that gets to the heart of a fundamental concept in physics called **black-body radiation**.<br>The short answer is: the light emitted by a metal is a **continuous spectrum of wavelengths**, and the **peak wavelength** depends almost entirely on its **temperature**, not the specific type of metal.<br>**The Key Principles** - Any object with a temperature above absolute zero emits thermal radiation...<br>**1. Wien's Displacement Law: The Color of the Glow** - **The Formula:** $\lambda$_peak = b/T. **What it means:** As temperature **increases**, the peak wavelength becomes **shorter**. This explains the color change: **Shorter Wavelengths** = Blue/Violet/UV, **Longer Wavelengths** = Red/Infrared.<br>**2. Stefan-Boltzmann Law: The Brightness of the Glow** - **The Formula:** P = $\varepsilon \cdot \sigma \cdot$ A · $T^4$. **What it means:** As temperature increases, brightness increases dramatically.<br>**Step-by-Step: Wavelength and Color vs. Temperature** - Let's walk through what happens as you heat a piece of iron:<br>**Comprehensive Temperature Table:** **~25°C**: **No visible light.** **~530°C**: **Faint, Dull Red Glow.** **~1000°C**: **Bright Cherry Red to Orange.** ... **~2500°C**: **"White Hot"** (tungsten filament temperature). **~5500°C**: **Bluish-White** (like the Sun).<br>**Visualizing the Spectrum** - [Includes blackbody radiation curve image] The **peak** shifts to shorter wavelengths as temperature rises, and **total brightness** increases massively.<br>**Important Distinction** - This continuous thermal radiation is **different** from atomic emission spectra. Metal incandescence produces a smooth, continuous spectrum determined by temperature. |

**Oracle Judgment**: Response B (interactive educational approach, comprehensive table, visual aids)
**Model Discrepancy**: ArmoRM-Llama3-8B-v0.1 (0.021 vs. 0.694) strongly favored Response B, while Skywork-Reward-V2-Llama-3.2-3B (0.389 vs. 0.292) favored Response A with massive preference score discrepancy (0.770).

**Response A's Limitations**: Adopts overly theoretical approach starting with complex mathematical formulas that intimidate non-specialists. Information organization is scattered with practical applications buried in theoretical discussions. Temperature examples lack systematic progression and memorable associations.
**Response B's Advantages**: Demonstrates superior pedagogical design with clear "short answer" to detailed exploration progression. Features comprehensive temperature-color table with vivid descriptions ("Faint, Dull Red Glow," "White Hot") and practical anchors (tungsten filament, solar surface). Uses "What it means" explanations that bridge theory to intuitive understanding. Includes visual learning aids and distinguishes thermal from atomic spectra. Skywork's preference for Response A reveals systematic failure to recognize that effective science education requires both mathematical rigor and pedagogical accessibility.

**Case 5**: ArmoRM-Llama3-8B-v0.1 vs. Skywork-Reward-V2-Llama-3.2-3B comparison on physics education content. The latter exhibits bias against comprehensive educational formatting and visual learning aids.