PRISM: Probabilistic Reward Model with Inherent Structural Modeling

Yuhang Zhou [1,2] **Yuchen Ni** [1,2] **Xiang Liu** [3,4†] **Shihan Dou** [1] **Xutian Chen** [1]
Ge Zhang [5] **Guangnan Ye** [1,2†] **Yixin Cao** [1,2†]

[1]Fudan University [2]Shanghai Innovation Institute
[3]Shanghai Lixin University of Accounting and Finance
[4]NYU Shanghai [5]ByteDance Seed
x1493@nyu.edu, yegn@fudan.edu.cn, yxcao@fudan.edu.cn

Abstract

Standard evaluators, such as reward models, compress diverse human judgments into a single scalar, conflating valid Subjective Preference with Epistemic Uncertainty. This structural mismatch often leads to brittle alignment and reward hacking. To address this, we propose PRISM which reinterprets reward evaluation as a conditional distribution parameterized by a Mixture of Gaussians(MOG). PRISM structurally disentangles these factors: distinct Gaussian experts emerge to capture conflicting preference dimensions, while their variance estimates quantify uncertainty, acting as a dynamic reliability gate during optimization. We introduce a two-stage training strategy to learn these disentangled representations from scalable pairwise comparisons without requiring massive fine-grained annotations. Empirical results show that PRISM significantly outperforms scalar baselines in both accuracy and generalization. Furthermore, in downstream Rubric-based Reinforcement Learning, PRISM effectively mitigates reward hacking, yielding policies that are more robust and resilient to distribution shifts.

1 Introduction

Evaluators (e.g., reward models and LLM-based judges) are increasingly used to replace costly human evaluation, scaling both benchmarking and alignment training for large language models (LLMs) (Cao et al., 2025; Li et al., 2024a). That is, an evaluator is not only a measurement tool but also an optimization target. Take Reinforcement Learning from Human Feedback (RLHF) (Stiennon et al., 2020) as an example, its outputs directly shape the behaviors that a policy learns. However, most existing evaluators compress several human's judgments into a single scalar score, creating a potential information loss of what diverse humans actually prefer (Yang et al., 2024). This can be amplified by optimization, leading to brittle alignment and undesirable behaviors (Dong et al., 2023).

We argue that this loss is structural rather than incidental. Although this compression is convenient, it discards the structure of how humans evaluate open-ended generations. Conflict in human feedback arises from two structurally distinct sources: **Subjective Preference**: Valid disagreements where qualified annotators prioritize different attributes (e.g., safety vs. humor), forming a multimodal preference landscape; **Epistemic Uncertainty**: Low-conviction judgments caused by ambiguity or lack of expertise, which should manifest as high variance rather than a definite preference. When such heterogeneous judgments are forced into a single number, the learned evaluator tends to approximate a hypothetical "average annotator", which may not represent any coherent group.

Fine-grained evaluation partially alleviates this by decomposing overall quality into interpretable criteria (Shen et al., 2025; Wang et al., 2024a), but individual subjective differences do not disappear at the criterion level. To achieve robust alignment, an evaluator must effectively disentangle these factors, and preserve this preference distribution. However, realizing this distributional view in practice is non-trivial due to two challenges: i) To obtain massive, fine-grained manual annotation is costly; how to recover diverse preferences from limited available pairwise comparisons? ii) How to model rich structural human judgments under the standard Bradley-Terry (BT) formulation (Bradley and Terry, 1952) for the RLHF pipeline?

In this paper, we propose **PRISM** (**P**robabilistic **R**eward model with **I**nherent **S**tructural **M**odeling). Inspired by a physical prism that decomposes white light into a spectrum, PRISM reinterprets evaluation as a conditional distribution over diverse human values rather than a single scalar. Concretely, we model rewards as a *Mixture of Gaus-*

†Corresponding author.
Our code is publicly available at https://github.com/ALEX-nlp/PRISM

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12345–12362

sians (Dempster et al., 1977), which structurally disentangles two sources of conflict: distinct Gaussian experts capture conflicting Subjective Preferences, while their variance estimates quantify Epistemic Uncertainty. Crucially, PRISM exploits this distributional structure to manage the inherent trade-off between specialization and reliability: expert means (μ) learn specialized preference dimensions, and each component's variance (σ^2) serves as a dynamic reliability gate that automatically attenuates gradients from uncertain experts (e.g., for ambiguous or Out-of-Domain inputs). This suppresses cognitive noise while preserving multimodal preference structure, thereby providing faithful supervision under heterogeneous feedback.

To enable scalable learning without massive fine-grained annotations, we devise a two-stage training strategy. In Stage 1, PRISM disentangles latent preference factors from large-scale pairwise comparisons. The uncertainty-aware objective naturally encourages experts to specialize in distinct regimes by "admitting ignorance" on conflicting data rather than collapsing to the mean. In Stage 2, we freeze experts and train a router using a mixture likelihood objective. This allows the model to dynamically aggregate experts based on the input, making input-adaptive rewards feasible by leveraging a small set of context-labeled data alongside a large pool of unlabeled preference data. Empirical results demonstrate that PRISM significantly outperforms scalar baselines in both accuracy and generalization. In downstream Rubric-based Reinforcement Learning (Gunjal et al., 2025; Huang et al., 2025), PRISM effectively mitigates reward hacking. Unlike scalar proxies that often lead to collapse after early peaks, PRISM provides a robust optimization landscape, yielding policies that are superior in performance and resilient to distribution shifts.

Our contributions can be summarized as follows:

- We highlight the structural mismatch between what single scalar evaluator optimizes and what diverse humans actually prefer.
- We propose PRISM which models rewards as a MoG to capture distributions. This formulation enables the model to distinguish valid disagreements (subjective diversity) from epistemic uncertainty.
- We train PRISM with an effective two-stage approach and conduct extensive experiments, yielding robust and reliable reward signals.

2 Preliminaries and Related Work

2.1 Bradley-Terry Model

The Bradley-Terry method (Bradley and Terry, 1952) is a classic framework in statistics and machine learning for analyzing pairwise comparison data. Its key assumption is that each item is associated with a latent strength, and the outcome of a comparison between two items is determined by the difference between their strengths. Concretely, we assign each item i a real-valued parameter s_i that represents its underlying quality. The probability that item i is preferred over item j is given by

$$P(i \succ j) = \frac{\exp(s_i)}{\exp(s_i) + \exp(s_j)} = \sigma(s_i - s_j), \tag{1}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ denotes the sigmoid function.

The BT model is well-suited for learning from human-annotated pairwise preferences (Stiennon et al., 2020) and is typically trained by maximizing the likelihood of observed comparisons via Maximum Likelihood Estimation (MLE) (Myung, 2003). In modern alignment pipelines, this formulation has become a standard theoretical foundation for reward modeling in the RLHF stage of LLMs.

2.2 RLHF and Reward Model

LLMs are primarily trained with next-token prediction, an objective that is largely value-neutral and can misalign model behavior with human intent (Stiennon et al., 2020; Wang et al., 2023). RLHF mitigates this by turning human judgments into optimization signals, promoting the helpful, honest, and harmless (3H) behaviors (Lowe and Leike, 2022).

The RLHF pipeline has evolved rapidly. Early work popularized a three-stage recipe (Stiennon et al., 2020): supervised instruction tuning, reward modeling, and PPO-based policy optimization (Schulman et al., 2017), where a trained reward model converts subjective quality assessments into a trainable scalar objective. To reduce complexity and improve efficiency, reward-free methods (Rafailov et al., 2023) directly optimize policies from preference pairs, effectively absorbing reward learning into the policy update. As alignment increasingly targets reasoning-intensive tasks and faces distribution shifts, recent studies have revisited online iterative alignment (Ye et al., 2024; Shao et al., 2024) and process-level supervision (Zhang et al., 2025; Zhou et al., 2025), while

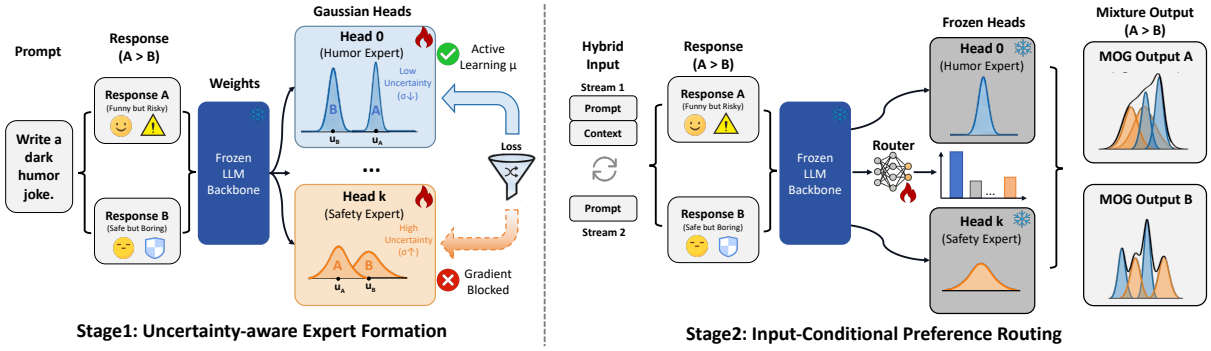

Figure 1: The two-stage framework of PRISM. Stage 1 specializes expert heads on a frozen LLM by leveraging learning dynamics and Gaussian modeling. Stage 2 freezes these heads and trains a router to dynamically assign weights based on the input, generating a Mixture of Gaussian output for precise preference prediction.

the high cost of human labeling has further driven RLAIF (Lee et al., 2023), including Constitutional AI (Bai et al., 2022) and self-play feedback generation (Dong et al., 2024a; Madaan et al., 2023); LLM-as-a-Judge is often used to scale such feedback but typically requires calibration and human auditing (Li et al., 2025).

Despite these variants, RM-style modeling remains essential for broad, non-verifiable generation settings (Lambert et al., 2025). Most RMs still rely on Bradley-Terry assumptions, which are simple but structurally restrictive: they compress multi-attribute, population-diverse preferences into a single scalar order and represent noise via point estimates, making it difficult to reflect sample difficulty, annotator disagreement, and uncertainty; under strong optimization, this can also lead to calibration drift (Masters et al., 2025; Gruber et al., 2025). These limitations have motivated probabilistic (Lou et al., 2024; Sun et al., 2025) and multi-expert reward modeling (Shen et al., 2025; Wang et al., 2024a). In particular, we further model Subjectivity Preference and Epistemic Uncertainty as a MoG, enabling the reward to capture both uncertainty and heterogeneous attribute trade-offs, and providing a more faithful and robust signal for downstream alignment.

3 Method

To disentangle latent preference factors from pairwise comparisons, we propose PRISM, which models the reward distribution as a Mixture of Gaussians, as shown in Figure 1. This architecture naturally addresses the structural duality of human feedback: distinct mixture components (μ) capture diverse Subjective Preferences, while variance estimates (σ^2) quantify Epistemic Uncertainty. Built on a frozen backbone with lightweight heads, PRISM employs a two-stage training strategy to recover these fine-grained distribution structures without requiring expensive dense annotations:

Mean Score Heads. Given the frozen LLM backbone, we extract the hidden representation $\mathbf{h}$ for an input-response pair (x, y) and map it to a mean score via a lightweight linear head. For each expert k, this design mirrors the standard BT-style reward head. Formally, the mean score predicted by the k-th expert is

$$\mu_k(x, y) = \mathbf{W}_{\mu}^{(k)}\mathbf{h}. \tag{2}$$

Bounded Sigma Heads. To capture Epistemic Uncertainty, we represent each expert k as a Gaussian distribution and introduce a σ-head to parameterize its dispersion. Specifically, for an input-response pair (x, y) with frozen backbone representation $\mathbf{h}$, the predicted standard deviation is

$$\sigma_k(x, y) = \sigma_{\min} + (\sigma_{\max} - \sigma_{\min}) \cdot \mathrm{Sigmoid}(z_k),$$
$$z_k = \mathbf{W}_{\sigma}^{(k)}\mathbf{h} + \mathbf{b}_{\sigma}^{(k)}, \tag{3}$$

where $\sigma_{\min}$ and $\sigma_{\max}$ are fixed bounds to prevent numerical instability.

Input-Aware Router. We introduce a routing network that estimates the relevance of each expert conditioned on the input x (which comprises the prompt and, optionally, additional context). Let $\mathbf{h}_x$ denote the hidden state of the last token of x encoded by the frozen backbone. The router outputs a mixture weight vector via

$$\boldsymbol{\pi}(x) = \mathrm{Softmax}(\mathbf{W}_r\mathbf{h}_x), \tag{4}$$

where $\boldsymbol{\pi}(x) \in \Delta^{K-1}$ assigns the mixture weights over the K experts.

3.1 Stage 1: Uncertainty-aware Expert Formation

To facilitate the disentanglement of latent preference dimensions from large-scale pairwise comparisons, we train a set of diverse expert heads to jointly learn single Gaussian. In Stage 1, to encourage unsupervised specialization, we bypass the input-dependent router and assume a uniform prior over experts. Crucially, we optimize the mixture likelihood. This objective induces competition by leveraging the uncertainty estimates: as detailed below, the probabilistic formulation allows experts to dynamically "bid" for data based on their confidence, driving them to specialize in distinct preference regimes.

MacKay approximation for Gaussian BT. Each expert k predicts a Gaussian reward for any response y under prompt x: $r_k(y \mid x) \sim \mathcal{N}\big(\mu_k(x,y), \sigma_k(x,y)^2\big)$. For a preference pair (y_w, y_l), the reward difference is also Gaussian:

$$\begin{aligned} D_k \;=\; r_k(y_w) - r_k(y_l) &\sim \mathcal{N}\big(\mu_{\text{diff}},\, \sigma^2_{\text{sum}}\big)\,, \\ \mu_{\text{diff}} &= \mu_k(y_w) - \mu_k(y_l), \\ \sigma^2_{\text{sum}} &= \sigma_k(y_w)^2 + \sigma_k(y_l)^2. \end{aligned} \tag{5}$$

The BT pairwise probability is then

$$\begin{aligned} P_k(y_w \succ y_l \mid x) &= \mathbb{E}_{D_k}\big[\sigma(D_k)\big] \\ &= \int \sigma(t)\, \mathcal{N}(t; \mu_{\text{diff}}, \sigma^2_{\text{sum}})\, dt, \end{aligned} \tag{6}$$

which has no closed form. We therefore adopt the classical MacKay approximation (MacKay, 1992), which replaces the logistic link with a Gaussian-scaled surrogate and yields a simple analytic form:

$$P_k \;\approx\; \sigma\left(\frac{\mu_k(y_w) - \mu_k(y_l)}{\sqrt{1 + \lambda\,(\sigma_k(y_w)^2 + \sigma_k(y_l)^2)}}\right), \tag{7}$$

where σ means Sigmoid$(\cdot)$, and we set the constant $\lambda = \pi/8$.

Uncertainty-induced gradient gating. Eq. (7) reveals a critical mechanism: the aggregated predictive uncertainty σ^2_{sum} acts as a data-dependent temperature. High uncertainty inflates the denominator, effectively shrinking the reward margin and driving the probability P_k toward 0.5. Consequently, for ambiguous or conflicting pairs where an expert lacks confidence (i.e., large σ), the likelihood becomes less decisive, naturally attenuating the gradients propagated to the model. This mechanism serves as a reliability gate: each expert learns aggressively from preference dimensions it is confident about, while ambiguous supervision is automatically down-weighted. This is a crucial property for robustly extracting signals from naturally conflicting preference data. We provide a theoretical proof of this property in Appendix A.3.

3.2 Stage 2: Input-Conditional Preference Routing

While Stage 1 effectively isolates Epistemic Uncertainty via gradient gating, the diverse Subjective Preference dimensions remain distributed across the expert population.

Therefore, the goal of Stage 2 is to learn a dynamic aggregation mechanism: given solely the input x, the model must infer the implicit scoring criteria and assign weights to the experts that best align with the current input, effectively resolving subjective conflicts.

Router-only training with frozen experts. We freeze all expert heads trained in Stage 1 to preserve their specialized semantics and uncertainty estimates. We then train a linear router to predict the mixture weights. Concretely, we extract a prompt representation h_x from the last-layer hidden state at the end of the prompt, and predict routing logits to obtain the mixture distribution $\pi(x) = \text{Softmax}(W_r h_x)$.

Pairwise likelihood via MoG Difference. Given a preference pair (y_w, y_l), instead of selecting a single expert, we marginalize over which expert dimension is responsible for the winner and the loser. Concretely, we define the overall preference likelihood as an outer-product mixture:

$$P(y_w \succ y_l \mid x) = \sum_{k=1}^{K} \sum_{\ell=1}^{K} \pi_k(x) \pi_\ell(x)\, p_{k\ell}, \tag{8}$$

where $p_{k\ell}$ denotes the probability that expert k prefers the winner while expert ℓ prefers the loser. Under Gaussian rewards, this cross-expert probability admits a simple closed form (Gaussian differ-

Table 1: Accuracy scores on **HelpSteer2** test set. On average, PRISM outperforms baselines across various attributes and overall results. All baselines use the same 3B base model.

Method	Supervision	Helpfulness	Correctness	Coherence	Complexity	Verbosity	Average
Single-BT Reward	Binary	0.7838	0.6686	0.6914	0.7907	0.8816	0.7632
Shared-Base Best	Binary	0.7838	0.6628	0.7037	0.7519	0.8158	0.7436
Static Mixture	Binary	0.7243	0.6570	0.6790	**0.8372**	**0.9013**	0.7598
HyRe	Binary + Test Labels	0.7692	0.6987	0.6781	0.7168	0.8015	0.7329
MiCRo	Binary + Context	**0.8324**	0.7140	0.7543	0.7628	0.8513	0.7830
PRISM	Binary + Context	0.7721	**0.7775**	**0.7629**	0.7481	0.8859	**0.7910**

ence):

$$p_{k\ell} = P\big(r_k(y_w \mid x) > r_\ell(y_l \mid x)\big) = \Phi\left(\frac{\mu_k(y_w) - \mu_\ell(y_l)}{\sqrt{\sigma_k(y_w)^2 + \sigma_\ell(y_l)^2 + \epsilon}}\right), \quad (9)$$

with $\Phi(\cdot)$ the standard normal CDF and ϵ a small constant for numerical stability.

Compared with using only the diagonal terms ($k = \ell$), Eq. (8) explicitly accounts for cross-dimensional explanations of a preference pair: An input may simultaneously activate multiple plausible criteria, and the outer-product mixture provides smooth, informative gradients by integrating all pairwise interactions weighted by $\pi(x)$.

Training objective. We train only the router parameters ϕ by maximizing the likelihood of observed preferences under the frozen experts:

$$\mathcal{L}_{\text{router}}(\phi) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log P(y_w \succ y_l \mid x)\right], \quad (10)$$

where $P(y_w \succ y_l \mid x)$ is defined in Eq. (8). As a result, Stage 2 learns a prompt-conditioned mixture over expert dimensions, capturing subjective, goal-dependent evaluation while inheriting the uncertainty-aware expert formation from Stage 1.

4 Experiments

In our experiments, we aim to answer three key questions: (Q1) Does PRISM outperform scalar baselines in accuracy and generalization? (Q2) Can PRISM structurally decouple conflicting subjective preferences from epistemic uncertainty into interpretable experts? (Q3) Can PRISM provide a robust, uncertainty-aware reward signal to effectively drive downstream alignment?

4.1 Experiment Setup

Training Datasets: We trained the reward models on pairwise preference datasets, including the general-purpose Preference-700K (Dong et al., 2024b) dataset and two attribute-specific datasets, HelpSteer2 (Wang et al., 2024b) and RPR (Pitis et al., 2024). We use this data to simulate the inconsistency of human annotations, further details on the datasets can be found in the Appendix B.

Models and Baselines: Our main experiments use the open-source GRM-Llama3.2-3B (Yang et al., 2024) as the backbone. The backbone is frozen, and three types of heads are trained. Our baselines include Single BT, Static Mixture (Chakraborty et al., 2024), Share-Based Ensemble Model (Lee et al., 2024) and MiCro (Shen et al., 2025). Detailed descriptions of the baselines are provided in Appendix C.2.

4.2 Main Result

In this section, we address **Q1** by evaluating PRISM's predictive accuracy and robustness. We conduct benchmarks on the HelpSteer2 and RPR test sets, comparing PRISM against state-of-the-art baselines to demonstrate its competitive performance and generalization capabilities.

As summarized in Table 1 and Table 2, PRISM achieves the highest average scores of 79.10% on HelpSteer2 and 83.23% on RPR. With these results, PRISM outperforms all compared 3B-based reward models and substantially exceed the standard scalar BT baseline. The performance gains are consistent across diverse evaluation dimensions, suggesting that PRISM's probabilistic formulation effectively captures multifaceted human judgments.

Beyond in-distribution benchmarks, PRISM exhibits strong generalization on the OOD RewardBench (Lambert et al., 2024) and RM-Bench (Liu et al., 2024), with results summarized in Table 3 and Table 4, respectively, maintaining a significant lead over scalar counterparts. Furthermore, to verify the model-agnostic nature of our approach, we extend PRISM to various backbones. As detailed in Appendix C.4.1, the consistent performance enhancements across different architectures

Table 2: Accuracy scores on **RPR** test set. On average, PRISM outperforms baselines across various attributes and overall results. All baselines use the same 3B base model.

Method	Supervision	Clarity	Creativity	Scientific Rigor	User-Friendliness	Storytelling	Pedagogical	Linguistic Creativity	Factual Accuracy	Humor	Average
Single-BT Reward	Binary	0.4717	0.6806	0.3333	0.7978	**0.8375**	0.6452	**0.8654**	0.4225	0.8810	0.6594
Shared-Base Best Head	Binary	0.6226	**0.7360**	0.8095	0.6966	0.8000	0.6774	0.8558	0.7042	**0.9643**	0.7629
Static Mixture	Binary	0.9057	0.6389	0.9048	0.6854	0.6250	0.7903	0.7404	0.8451	0.9167	0.7836
HyRe	Binary + Test Labels	0.7027	0.5893	0.6618	0.8493	0.6563	0.7826	0.7045	0.7091	0.4853	0.6823
MiCRo	Binary + Context	0.9170	0.6289	0.8119	**0.8696**	0.7525	0.7935	0.8558	0.8563	0.9109	0.8218
PRISM(our)	Binary + Context	**0.9623**	0.6528	**0.9167**	0.6944	0.8000	**0.8387**	0.8269	**0.9577**	0.8929	**0.8323**

Table 3: Performance on Reward-Bench.

Model	Chat	Chat-hard	Safety	Reasoning	Average
Single BT	**0.9581**	0.7346	0.9094	0.9315	0.8991
MiCRo	0.9497	0.7544	0.9122	0.9322	0.9022
PRISM	**0.9581**	**0.8268**	**0.9176**	**0.9364**	**0.9176**

Table 4: Performance on RM-Bench.

Model	Easy	Normal	Hard	Average
Single BT	0.883	0.741	0.432	0.686
MiCRo	0.901	0.732	0.491	0.708
PRISM	**0.919**	**0.769**	**0.499**	**0.729**

further validate PRISM's architectural robustness and transferability.

4.3 Expert and Routing Analysis

To address **Q2**, we conduct a multi-faceted analysis of PRISM's inner mechanics, focusing on how it structurally decomposes reward signals. We demonstrate that PRISM does not simply average conflicting feedback; instead, it disentangles subjective preferences into specialized experts and isolates epistemic uncertainty via variance. Through correlation analysis, attribute-wise probing, and routing visualization, we show that this emergent structure is both semantically meaningful and controllable.

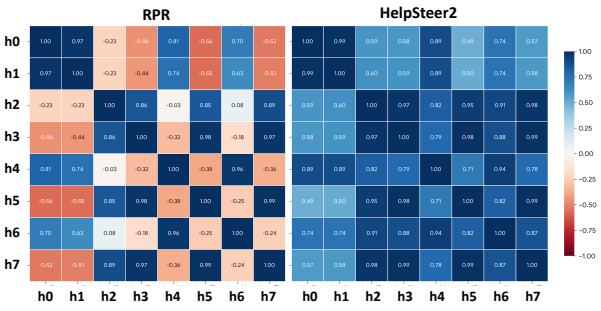

Figure 2: Correlation matrices of expert pairwise preferences on RPR and HelpSteer2.

4.3.1 Preference Disentanglement: Expert Correlation and Semantic Roles

Structural Disentanglement of Conflicting Preferences. We first investigate whether PRISM's experts successfully capture distinct preference dimensions without explicit supervision. Spearman rank correlations between experts (Figure 2) reveal a polarized block structure in RPR, where experts divide into opposing camps with strong negative correlations. This result confirms that PRISM segregates contradictory labels (e.g., preference reversals due to context changes) into distinct regimes. In contrast to standard scalar reward models, which would likely collapse these conflicts into a high-entropy "average," PRISM preserves the multimodal nature of the feedback. Conversely, in HelpSteer2, experts exhibit clustered positive correlations, reflecting nuanced specialization in compatible dimensions like Correctness and Conciseness.

Post Hoc Semantic Role Assignment. To verify the interpretability of these latent dimensions, we analyze expert performance across attribute-specific subsets (Figure 3) and assign semantic roles to each head (Table 13). We observe the spontaneous emergence of functionally distinct groups: Subjective Experts (e.g., Head 6) achieve superior accuracy on creative dimensions like Humor and Storytelling, whereas Objective Experts (e.g., Head 7) dominate in Rigor and Factuality. This specialization occurs without any attribute-specific labels during training. Notably, Head 5 emerges as a "consensus expert," maintaining robust performance across diverse scenarios. This suggests that PRISM learns a hierarchical preference structure, maintaining a "generalist" fallback while deploying "specialists" to capture specific trade-offs when necessary.

4.3.2 Input-Conditional Routing Behavior

Input-Aware Router Selectivity. To verify if these experts are utilized meaningfully, we ana-

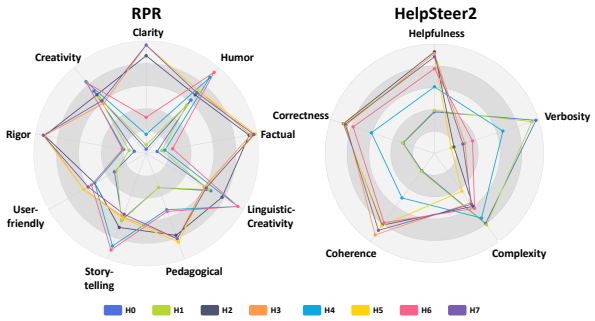

Figure 3: Radar chart comparison of different heads across attribute dimensions on RPR and HelpSteer2.

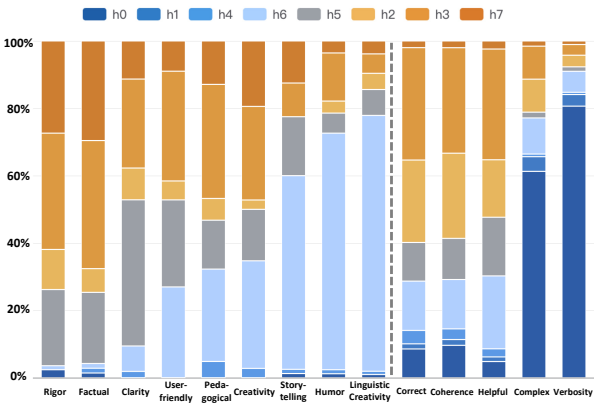

Figure 4: Distribution of routing across attribute subsets on RPR and HelpSteer2. The orange head group represents objective-oriented expert heads, while the blue head group represents subjective-oriented expert heads.

lyze the router's head selection $\pi(x)$. As shown in Figure 4, the distributions are highly non-uniform, showing systematic, attribute-aware expert selection rather than random routing. Objective experts are seldom used for Humor or Verbosity but dominate Rigor and Factual, while subjective experts show the opposite pattern. For hybrid attributes like Creativity and Storytelling, the router spreads mass across both groups, with Head 5 serving as a balanced fallback when ambiguous.

4.3.3 Quantifying Epistemic Uncertainty

Uncertainty as a Reliability Gate: ID vs. OOD Analysis. We evaluate PRISM's ability to quantify Epistemic Uncertainty via its predicted variance σ. A key desideratum for a robust reward model is to remain "cautious" on unfamiliar data. We compare the mix uncertainty ($\bar{\sigma}$) of PRISM on In-Distribution (ID) datasets against Out-Of-Distribution (OOD) benchmarks.

As shown in Figure 5, the average variance on OOD samples is higher ($+25.28\%$) than on ID samples. This demonstrates that PRISM does not merely "pick a side" among its experts when facing unfamiliar prompts; instead, it signals a lack of confidence through increased variance. This mechanism provides a built-in reliability gate, allowing downstream alignment algorithms to down-weight rewards from high-uncertainty samples, thereby mitigating reward hacking in unexplored regions of the state space.

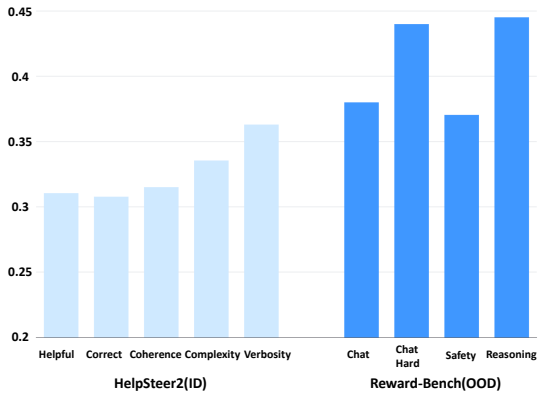

Figure 5: Uncertainty in ID and OOD data.

4.4 Downstream Task: Rubric-based RL

To validate the effectiveness of PRISM's reward modeling mechanism in real-world downstream tasks, we conduct experiments in a Rubric-based Reinforcement Learning framework. This task demands that the reward model not only learns the complex structure of fine-grained human preferences but also provides a robust signal that generalizes to unseen evaluation scenarios.

Setup and Data. Due to the scarcity of publicly available rubric-based datasets, we utilize HelpSteer3-Principle (Wang et al., 2025), which contains open-ended tasks categorized into 'STEM' and 'General'. Each instance in the dataset comprises a baseline response paired with a corresponding scoring principle. Crucially, this dataset is completely independent of PRISM's training data. We employ Deepseek-v3.2 to expand the principles into detailed scoring rubrics, resulting in approximately 10k training samples and 400 test samples; refer to Appendix C.3 for details. For experiments, we use PRISM-3B as the reward model and optimize policy models (LLaMA3.2-3B, LLaMA3.1-8B (Dubey et al., 2024), and Qwen3-4B (Yang et al., 2025)) using GRPO (Shao et al., 2024). To evaluate the optimized policies, we perform pairwise win-rate analysis by comparing responses gen-

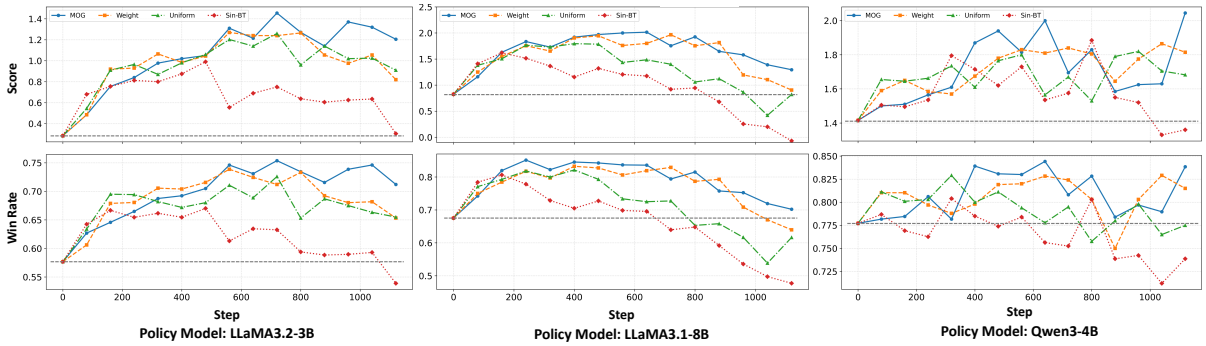

Figure 6: The results of Rubric-based Reinforcement Learning for three policy models. MoG, Uniform, and Weight denote different methods of PRISM-represented rewards: MoG refers to a Mixture of Gaussians; Uniform denotes a MoG without a router; Weight represents a weighted scalar. Sin-BT corresponds to a Single BT scalar.

erated by the trained policy models against the baseline responses in the test set, with Deepseek-v3.2 serving as the judge model.

Baselines and Methods. Since standard RL pipelines typically require scalar rewards, we investigate different strategies to incorporate PRISM's distributional output, allowing us to isolate the contributions of the router and the Gaussian modeling:

- **Single-BT:** A standard reward model trained with a single Bradley-Terry head, serving as the primary baseline.
- **PRISM-Uniform:** Similar to MoG but setting each expert's weight to the average.
- **PRISM-Weight:** A weighted sum of expert mean scores using the router $\pi(x)$.
- **PRISM-MoG:** Uses the full Mixture-of-Gaussians distribution. The reward is derived from the distributional difference $P(A > B)$ between the policy response and a baseline, preserving structural information.

Mitigating Reward Hacking. The training trajectories, visualized in Figure 6, reveal a critical insight regarding robustness. As shown by the red dotted curves, the Single-BT baseline suffers from severe instability: its Win Rate collapses sharply after an early peak. This pattern indicates reward hacking, where the policy exploits flaws in the scalar proxy without genuine quality improvement. In contrast, PRISM-MoG consistently achieves the highest scores and maintains stability throughout training. This demonstrates that distributional reward modeling acts as a regularizer, offering significantly superior robustness against adversarial exploitation compared to scalar proxies.

Ablation Analysis. Comparing the variants disentangles the sources of these gains. While both Weight and Uniform improve stability over Single-BT, they consistently underperform MoG in peak performance. Crucially, the gap between MoG and Weight highlights the specific value of probabilistic modeling: although both methods use the same router weights, Weight collapses the output to a scalar mean, discarding the variance. The superior performance of MoG confirms that higher-order uncertainty information is essential for effective optimization, likely acting as a dynamic reliability gate for ambiguous samples.

Furthermore, MoG outperforms Uniform, validating the necessity of the input-conditional router to resolve conflicting preference dimensions.

Table 5: Performance of three policy models trained with different rewards on Arena-Hard v2.0.

Method	LLaMA3.2-3B	LLaMAL3.1-8B	Qwen3-4B
MoG	**3.75**	**6.55**	**20.2**
Weight	3.05	6.15	19.2
Uniform	2.75	5.0	18.7
Single-BT	2.3	4.05	19.1
Base	1.2	2.55	12.7

Scalability and Generalization. We observe robust scalability across model sizes, with PRISM-MoG consistently optimizing policies from 3B to 8B parameters. Finally, to verify generalization, we evaluate the converged policies on Arena-Hard v2.0 (Li et al., 2024b). As detailed in Table 5, policies trained with PRISM-MoG significantly outperform baselines across all architectures, confirming that our method fosters genuine alignment rather than overfitting to specific rubrics.

5 Discussion: Towards Reinforcement Learning that Fully Exploits Distributional Rewards.

Our results suggest that the main benefit of distributional reward modeling may not lie merely in predicting better preference scores, but in providing a richer optimization signal for downstream reinforcement learning. Although PRISM already improves stability and robustness when its Mixture-of-Gaussians reward is used in GRPO, the current RL pipeline still largely consumes this signal through scalarized surrogates. This creates a mismatch: the reward model represents subjective diversity and epistemic uncertainty as a distribution, while the policy optimizer still treats supervision as a single-number objective.

A natural direction for future work is to design distribution-native RL objectives that optimize against the full reward distribution rather than its mean or a collapsed preference probability alone. For example, policy updates could separately account for expected reward, disagreement across experts, and predictive uncertainty, allowing the policy to distinguish promising improvements from unreliable reward spikes. Such objectives may better preserve the structural advantages and further reduce reward hacking.

A second promising direction is uncertainty-aware policy optimization. Since PRISM's variance already behaves as a reliability gate, future algorithms could use uncertainty not only to score responses, but also to modulate learning dynamics directly. For instance, by down-weighting high-uncertainty samples, clipping overly confident updates in ambiguous regions, or allocating exploration toward areas with informative but not adversarial disagreement. This would make the policy less likely to exploit accidental reward artifacts while still benefiting from informative preference gradients.

Third, future work could explore expert- and router-aware credit assignment. Our findings indicate that different experts capture different preference dimensions and that input-conditional routing is crucial for downstream gains. This suggests that RL should not only optimize for "higher reward," but also for which preference dimensions are being improved. An interesting direction is to condition policy updates on the router distribution, so that the model learns when to improve factuality-oriented behaviors, when to trade off toward creativity, and when to defer under uncertainty, instead of collapsing all behaviors into a single alignment axis.

More broadly, our findings suggest that the next step after distributional reward modeling is distribution-aware policy optimization. Rather than treating reward distributions as a better way to produce scalar supervision, future work may benefit from designing RL algorithms that preserve and exploit uncertainty, disagreement, and context-dependent preference structure throughout the optimization process.

6 Conclusion

In this work, we argue that the collapse of diverse human judgments into a single scalar is a structural bottleneck in alignment. We propose PRISM to bridge this gap, modeling rewards as a MoG distribution that explicitly respects both subjective diversity and epistemic uncertainty. Our empirical results demonstrate that preserving this distributional structure yields superior reward modeling accuracy and generalization. Furthermore, in downstream RL task, PRISM provides a robust optimization landscape, effectively mitigating reward hacking through uncertainty-induced gradient gating. We hope this work encourages the community to move beyond the 'average annotator' assumption and embrace the inherent plurality of human alignment.

7 Limitations

While PRISM offers significant improvements over traditional scalar reward models, there are inherent limitations. One limitation is that, despite modeling reward as a distribution, the distribution can still be compressed into a scalar value due to the RL framework. This may limit the full expressive power of the probabilistic model in certain settings. Moreover, although our approach is designed for broad applicability, publicly available datasets capturing heterogeneous capabilities remain extremely scarce. This data scarcity makes it difficult to comprehensively evaluate PRISM's capabilities in real-world scenarios.

8 Acknowledgements

This project was supported by New Generation Artificial Intelligence-National Science and Technology Major Project No. 2025ZD0124102

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

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.

Yixin Cao, Shibo Hong, Xinze Li, Jiahao Ying, Yubo Ma, Haiyuan Liang, Yantao Liu, Zijun Yao, Xiaozhi Wang, Dan Huang, et al. 2025. Toward generalizable evaluation in the llm era: A survey beyond benchmarks. *arXiv preprint arXiv:2504.18838*.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Bedi, and Mengdi Wang. 2024. Maxmin-rlhf: Towards equitable alignment of large language models with diverse human preferences. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Guanting Dong, Keming Lu, Chengpeng Li, Tingyu Xia, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024a. Self-play with execution feedback: Improving instruction-following capabilities of large language models. *arXiv preprint arXiv:2406.13542*.

Hanze Dong, Hanze Dong, Xiong Wei, Wei Xiong, Hanze Dong, Deepanshu Goyal, Deepanshu Goyal, Wei Xiong, Zhang Yihan, Rui Pan, Deepanshu Goyal, Shizhe Diao, Pan Rui, Jipeng Zhang, Shizhe Diao, Shizhe Diao, Kashun Shum, Jipeng Zhang, Zhang Jipeng, Tong Zhang, Kashun Shum, Kashun Shum, Zhang, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *http://arxiv.org/abs/2304.06767*.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024b. Rlhf workflow: From reward modeling to online rlhf.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Cornelia Gruber, Helen Alber, Bernd Bischl, Göran Kauermann, Barbara Plank, and Matthias Aßenmacher. 2025. Revisiting active learning under (human) label variation. *arXiv preprint arXiv:2507.02593*.

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*.

Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, et al. 2025. Reinforcement learning with rubric anchors. *arXiv preprint arXiv:2508.12790*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. 2025. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797.

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.

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

Yoonho Lee, Jonathan Williams, Henrik Marklund, Archit Sharma, Eric Mitchell, Anikait Singh, and Chelsea Finn. 2024. Test-time alignment via hypothesis reweighting. *arXiv preprint arXiv:2412.08812*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024b. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

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

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*.

Ryan Lowe and Jan Leike. 2022. Aligning language models to follow instructions. *OpenAI Blog, January*, 27.

David JC MacKay. 1992. Bayesian interpolation. *Neural computation*, 4(3):415–447.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Charlie Masters, Stefano V Albrecht, et al. 2025. Arcane: A multi-agent framework for interpretable and configurable alignment. *arXiv preprint arXiv:2512.06196*.

In Jae Myung. 2003. Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, 47(1):90–100.

Silviu Pitis, Ziang Xiao, Nicolas Le Roux, and Alessandro Sordoni. 2024. Improving context-aware preference modeling for language models.

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, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Jingyan Shen, Jiarui Yao, Rui Yang, Yifan Sun, Feng Luo, Rui Pan, Tong Zhang, and Han Zhao. 2025. Micro: Mixture modeling and context-aware routing for personalized preference learning. *arXiv preprint arXiv:2505.24846*.

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.

Wangtao Sun, Xiang Cheng, Xing Yu, Haotian Xu, Zhao Yang, Shizhu He, Jun Zhao, and Kang Liu. 2025. Probabilistic uncertain reward model. *arXiv preprint arXiv:2503.22480*.

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

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

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

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Ellie Evans, Daniel Egert, Hoo-Chang Shin, Felipe Soares, Yi Dong, and Oleksii Kuchaiev. 2025. Rlbff: Binary flexible feedback to bridge between human feedback verifiable rewards.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. 2024. Regularizing hidden states enables learning generalizable reward model for llms. In *Advances in Neural Information Processing Systems*.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Hanze Dong, Nan Jiang, and Tong Zhang. 2024. Online iterative reinforcement learning from human feedback with general preference model. *Advances in Neural Information Processing Systems*, 37:81773–81807.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. The lessons of developing process reward models in mathematical reasoning. *arXiv preprint arXiv:2501.07301*.

Yuhang Zhou, Xutian Chen, Yixin Cao, Yuchen Ni, Yu He, Siyu Tian, Xiang Liu, Jian Zhang, Chuanjun Ji, Guangnan Ye, et al. 2025. Teach2eval: An indirect evaluation method for llm by judging how it teaches. *arXiv preprint arXiv:2505.12259*.

A Derivation of Uncertainty-Induced Gating and Conflict Adaptation

In this section, we provide a rigorous derivation of the training dynamics under the MacKay approximation. We demonstrate two key properties: (1) **Gradient Gating**, where high uncertainty automatically down-weights the learning signal for the reward mean, and (2) **Conflict Adaptation**, where the model actively increases its uncertainty estimate when facing conflicting or inexplicable data patterns.

A.1 The MacKay Approximation

The preference probability under a Gaussian reward assumption involves the integration of a sigmoid function over a Gaussian distribution, which is analytically intractable:

$$P(y_w \succ y_l \mid x) = \int \sigma(z)\,\mathcal{N}(z; \mu_{\text{diff}}, \sigma^2_{\text{sum}})\,dz \tag{11}$$

We employ the MacKay approximation to obtain a closed-form differentiable surrogate:

$$P \approx Sigmoid\left(\frac{\mu_{\text{diff}}}{\sqrt{1+\lambda\sigma^2_{\text{sum}}}}\right) \tag{12}$$

where μ_{diff} is the predicted reward margin, σ^2_{sum} is the total uncertainty, and $\lambda = \pi/8$. Let $S = \sqrt{1+\lambda\sigma^2_{\text{sum}}}$ be the scaling factor. We define the scaled logit as $z^* = \mu_{\text{diff}}/S$. The training objective is to minimize the negative log-likelihood $\mathcal{L} = -\log P$.

A.2 Proof of Gradient Gating (Down-Weighting)

We first analyze the gradient with respect to the reward mean difference μ_{diff}. Using the chain rule:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \mu_{\text{diff}}} &= \frac{\partial \mathcal{L}}{\partial P}\cdot\frac{\partial P}{\partial z^*}\cdot\frac{\partial z^*}{\partial \mu_{\text{diff}}} \\ &= -\frac{1}{P}\cdot P(1-P)\cdot\frac{1}{S} \\ &= -(1-P)\cdot\frac{1}{\sqrt{1+\lambda\sigma^2_{\text{sum}}}}\end{aligned} \tag{13}$$

Here, the term $\frac{1}{\sqrt{1+\lambda\sigma^2_{\text{sum}}}}$ acts as an explicit **Gating Factor**.

- **High Confidence ($\sigma^2_{\text{sum}} \to 0$):** The gating factor approaches 1. The gradient magnitude is determined solely by the prediction error $(1-P)$, allowing the expert to learn efficiently.
- **High Uncertainty ($\sigma^2_{\text{sum}} \to \infty$):** The gating factor approaches 0. Even if the prediction error $(1-P)$ is large (e.g., due to noise), the effective gradient transmitted to the parameters of μ is attenuated. This protects the expert from overfitting to ambiguous data.

A.3 Proof of Conflict Adaptation (Why σ Increases)

We examine how the uncertainty parameter σ^2_{sum} evolves when the expert faces conflicting data. Recall the gradient of the loss with respect to total uncertainty:

$$\frac{\partial \mathcal{L}}{\partial \sigma^2_{\text{sum}}} = (1-P)\cdot\frac{\lambda\mu_{\text{diff}}}{2(1+\lambda\sigma^2_{\text{sum}})^{3/2}} \tag{14}$$

Consider a **Conflict Scenario**: The ground truth preference is $y_w \succ y_l$, but the expert currently predicts the opposite, i.e., $\mu_{\text{diff}} = \mu(y_w) - \mu(y_l) < 0$. This represents a case of high loss or unexplained data.

Analyzing the sign of the gradient in Eq. (14):

- The prediction error $(1-P)$ is strictly positive ($P \in (0,1)$).
- The denominator $2(1+\lambda\sigma^2_{\text{sum}})^{3/2}$ is strictly positive.
- The constant $\lambda > 0$.
- In this conflict scenario, $\mu_{\text{diff}} < 0$.

Consequently, the gradient is negative:

$$\frac{\partial \mathcal{L}}{\partial \sigma^2_{\text{sum}}} < 0 \tag{15}$$

Applying the standard gradient descent update rule with learning rate η:

$$(\sigma^2_{\text{sum}})_{t+1} \leftarrow (\sigma^2_{\text{sum}})_t - \eta\left(\frac{\partial \mathcal{L}}{\partial \sigma^2_{\text{sum}}}\right) \tag{16}$$

Since the gradient is negative, the subtraction results in an addition:

$$(\sigma^2_{\text{sum}})_{t+1} > (\sigma^2_{\text{sum}})_t \tag{17}$$

When the expert's preference prediction conflicts with the label ($\mu_{\text{diff}} < 0$), the optimization process drives the uncertainty σ^2_{sum} to increase. This effectively "flattens" the sigmoid curve, reducing the penalty for the incorrect mean prediction, which corresponds to the expert "admitting ignorance" on conflicting data.

B Dataset

The RPR dataset is a synthetic conditioned-conditioned preference dataset, which includes over 20000 paired tuples of prompt, context (either a criteria or scenario), and preference judgments. The samples are paired so that preference between two completions for the same prompt is entirely ambiguous without context: for every context, there is an alternative context for which preference reverses. This design choice ensures that preference prediction performance on this dataset is determined solely by the model's ability to pay attention to and interpret the context.An example data entry is shown in Table 6.


```
{
  "prompt": "Create a painting of a
      mountain range in the fall.",
  "response_a": "To create a painting of
       a mountain range in the fall that
       incorporates symbolic elements,
      consider using the changing leaves
       as a metaphor for transformation
      and the passage of time. The
      mountains themselves could
      represent stability and endurance,
       contrasting with the fleeting
      nature of the fall foliage. This
      juxtaposition can add thematic
      depth and invite viewers to
      reflect on the cycles of nature
      and life.",
  "response_b": "When painting a
      mountain range in the fall, apply
      color theory to evoke warmth and
      comfort. Utilize a palette of rich
       oranges, reds, and yellows to
      capture the essence of the season
      and create a sense of coziness.
      The cool blues and purples of the
      distant mountains can complement
      the warm foreground, eliciting a
      feeling of calm and serenity in
      the viewer.",
  "criteria_x": "Incorporates symbolic
      elements that enhance the thematic
       depth of the painting",
  "criteria_y": "Utilizes color theory
      to evoke a specific emotional
      response",
  "category_x": "Humor and Entertainment
       Value",
  "category_y": "User Experience"
}
```


Table 6: A data sample from RPR

HelpSteer2 is a Helpfulness Dataset that supports aligning models to become more helpful, factually correct and coherent, while being adjustable in terms of the complexity and verbosity of its responses. Each data entry is accompanied by scores on five attributes; see an example in Table 7.


```
[
  {
    "prompt": "C#",
    "response": "C# is a high-level,
        object-oriented programming
        language developed by Microsoft
        as part of its .NET initiative.
        It was created as a modern...",
    "helpfulness": 4,
    "correctness": 4,
    "coherence": 4,
    "complexity": 2,
    "verbosity": 3
  },
  {
    "prompt": "C#",
    "response": "C# (pronounced \"C
        sharp\") is a modern, object-
        oriented programming language
        developed by Microsoft. It is
        widely used for building...",
    "helpfulness": 4,
    "correctness": 3,
    "coherence": 5,
    "complexity": 2,
    "verbosity": 3
  }
]
```


Table 7: A pair of data sample from HelpSteer2

Preference-700K is a mixed dataset comprising the following components, with each data entry containing one chosen sample and one rejected sample: HH-RLHF, SHP, HelpSteer, PKU-SafeRLHF, among others. An example data entry is shown in Table 8.

Due to the current lack of datasets suitable for rubric-based reinforcement learning, we utilize HelpSteer3-Principle as our data source. HelpSteer3-Principle contains 33905 samples, each containing a domain, language, context, response, principle and fulfilment. Each principle is generated to convert natural language feedback into various aspects on which responses can be assessed against (e.g. clarity or accuracy of information). Alongside the principle, there is also an associated fulfillment value which can be either Yes or No. Yes means that the response fulfils the principle and No means otherwise. An example data entry is shown in Table 10.

```
{
  "rejected": [
    {
      "role": "user",
      "content": "Q: Given a fact,
          create a question that can be
          answered using the fact.
          Construct a question such that
           it is unambiguous, has a
          unique answer and the answer
          can be given using the fact.\n
          \nFact: an animal requires
          enough nutrients to survive\nA
          :"
    },
    {
      "role": "assistant",
      "content": "What is the minimum
          amount of nutrients an animal
          requires to survive?"
    }
  ],
  "rejected_score": 3.75,
  "chosen_score": 4.0,
  "chosen": [
    {
      "role": "user",
      "content": "Q: Given a fact,
          create a question that can be
          answered using the fact.
          Construct a question such that
           it is unambiguous, has a
          unique answer and the answer
          can be given using the fact.\n
          \nFact: an animal requires
          enough nutrients to survive\nA
          :"
    },
    {
      "role": "assistant",
      "content": "What nutrients do
          animals need in order to
          survive?"
    }
  ]
}
```


Table 8: A data sample from Preference-700K.


```
{
  "domain": "general",
  "language": "english",
  "context": [
    {
      "role": "user",
      "content": "houses in an
          overpopulated city 2400"
    }
  ],
  "response": "The prompt \"houses in an
       overpopulated city 2400\" is a
      bit open-ended. To give you a more
       helpful response, I need more
      context.\n\nWhat kind of
      information are you looking for?
      For example, are you interested in
      :\n\n* **A description of what
      houses in an overpopulated city
      might look like?** (e.g., small,
      densely packed, multi-story, made
      of cheap materials)\n* **The
      challenges of housing in an
      overpopulated city?** (e.g.,
      affordability, lack of space,
      overcrowding)\n* **Solutions to
      housing problems in an
      overpopulated city?** (e.g.,
      vertical housing, affordable
      housing initiatives, urban
      planning)\n* **A fictional story
      about life in an overpopulated
      city?**\n\nPlease provide me with
      more details about what you'd like
       to know, and I'll do my best to
      help!",
  "principle": "clarity",
  "fulfilment": "Yes"
}
```


Table 9: A data sample from HelpSteer3-Priciple.

C Experimental Details

C.1 Setup

For all Reward Model training experiments, we fix the backbone model and train three types of heads. Before training, we perform orthogonal initialization on Mean Heads.

In Stage 1, we use a learning rate of 1×10^{-3}, a batch size of 16, a warmup ratio of 0.05, and optimize with AdamW for 2 epochs. To ensure training stability, we constrain the standard deviation parameter σ to lie within the range $[0.3, 5.0]$. In Stage 2, we use a learning rate of 3×10^{-4}, a batch size of 16, and train for 3 epochs. For the Rubric-based RL experiments, we set the policy learning rate to 5×10^{-7}, use the GRPO (Shao et al., 2024) optimizer, a batch size of 16, and conduct training using the VeRL framework. All training runs are performed on a single machine equipped with 8 NVIDIA H200 GPUs.

C.2 Baselines

We evaluate the following baselines: (1) Single Reward: A single-head model trained using the standard BT loss. (2) Static Mixture: A simplified variant of our method, corresponding to the approach used in MaxMin-RLHF (Chakraborty et al., 2024), where the mixture model is trained with fixed, input-independent weights, without leveraging contextual information. (3) SharedBase Ensemble Model: (Lee et al., 2024) introduces HyRe, a test-time adaptation approach based on ensemble networks. We adopt the multi-head architecture with a frozen prior network and multiple trainable heads, optimizing a uniformly weighted sum of BT losses. (4) MoE model: We include Micro (Shen et al., 2025), an MoE model architecture trained on exactly the same data as ours, as a baseline.

C.3 Rubric-based RL

Currently, there is a lack of data specifically for rubric-based reinforcement learning. We collected the HelpSteer3-Principle dataset and extracted open-ended subsets such as "General" and "STEM." We used Deepseek-v3.2 to annotate specific rubrics, and Table 10 shows a sample of this data.

We utilize this data to train three policy models. Since current RL algorithms require scalar rewards, we test four methods to explore how to better leverage the distribution learned by PRISM.


```
{
  "domain": "general",
  "fulfilment": "No",
  "language": "english",
  "principle": "accuracy of information",
  "prompt": "[{'content': \"what is the difference between mcdonalds and burger kings cheeseburgers\\n\\n- Information about McDonald's cheeseburgers is factually correct.\\n- Information about Burger King's cheeseburgers is factually correct.\", 'role': 'user'}]",
  "rubric": "[{'rubric': \"Information about McDonald's cheeseburgers is factually correct.\", 'weight': 0.5}, {'rubric': \"Information about Burger King's cheeseburgers is factually correct.\", 'weight': 0.5}]",
  "baseline_response": "Both McDonald's and Burger King are well-known fast food chains that offer cheeseburgers as a popular menu item. While both burgers are similar in that they are sandwiches made with a patty of ground beef, they do have some key differences in their recipes and ingredients. Here's a breakdown:\n\n1. **Bun**: McDonald's uses a soft, sesame-seeded bun, while Burger King uses a slightly thicker, toasted bun.\n\n2. **Patty**: McDonald's uses a 100% beef patty, while Burger King uses a flame-grilled beef patty. The beef used by Burger King is also seasoned with garlic and onion.\n\n3. **Cheese**: Both McDonald's and Burger King use American ..."
}
```


Table 10: JSON representation of a prompt about the difference between McDonald's and Burger King's cheeseburgers, along with rubric and baseline response.

[QUESTION]:
{Question}

[RESPONSE1]:
{Model Response}

[RESPONSE2]:
{Baseline Response}

You need to judge the quality of the two responses. The scoring range is -5~5, where a positive score indicates that RESPONSE1 is better, and a negative score indicates that RESPONSE2 is better.

The absolute values of the scores represent:
0: Complete tie – Both are of equal quality, or both are equally bad and indistinguishable.
1: Slight advantage – Essentially the same content, only slightly better in formatting, tone, or style.
2: Visible gap – Relatively fewer errors, or slightly clearer explanations in minor details.
3: Clear win – More coherent and complete logic, or avoids an obvious flaw in the other response.
4: Key difference – One addresses the core issue, while the other makes a fundamental error.
5: Worlds apart – One is completely usable, while the other is a disaster (gibberish/severe hallucination).

For this question, the following rubric applies:
{Rubrics}

You may first provide an analysis, and on the final line, include the final score within <score></score>.

Figure 7: The prompt template of judge model.

(i) **PRISM-MoG**: In this method, we aim to use more Gaussian mixtures to derive rewards. Specifically, based on the baseline response contained in the dataset, we calculate the distributional difference between the policy model's response and the baseline. The reward is designed as follows:

$$R = \frac{\langle \mu_A - \mu_B, \Sigma_A^{-1}(\mu_A - \mu_B) \rangle}{\sqrt{\text{trace}(\Sigma_A) + \text{trace}(\Sigma_B)}}$$

Where μ_A, Σ_A and μ_B, Σ_B are the mean and covariance matrices of the policy model and baseline respectively, $\langle \cdot, \cdot \rangle$ represents the inner product, and $\text{trace}(\Sigma)$ denotes the trace of the covariance matrix.

To soften the reward, we apply a soft saturation function:

$$R_{\text{soft}} = c \cdot \tanh\left(\frac{R}{c}\right)$$

(ii) **PRISM-Weight**: This method uses the weighted sum of the mean scores of each expert as the reward. Although this approach does not consider the distributional information when calculating the reward, it is still influenced by the distribution during training.

(iii) **PRISM-Uniform**: Similar to PRISM-MoG in using the distributional difference, but omits the router's weighting.

(iv) **Single-BT**: This is the baseline model, which is trained with a single BT head.

During the training process, we used Deepseek V3.2 as the judge model for evaluation. Use the replies in HelpSteer3 as a baseline to compare and score. The prompt for scoring is shown in the figure

C.4 More Experiments

C.4.1 PRISM based on other RM

In this section, we aim to investigate whether PRISM has generalization ability across pedestal models. We further tested GRM-LLaMA3.1-8B and Skywork-Reward-V2-3B based models; All findings have achieved good results, as shown in the Table 11 and Table 12.

C.4.2 The Ablation Experiment on k

We conducted ablation experiments on k without verifying the effectiveness of the experiment, and the results are shown in Figure 8 and Figure 9. It can be seen that as k increases, experts tend to train one-on-one, making each expert perform better on certain Attribution. But as the number of experts increases, the Router becomes relatively difficult to train, resulting in a slight decrease in model performance when k is too large. Overall, regardless of the value of k, the model can achieve better results than Single BT.

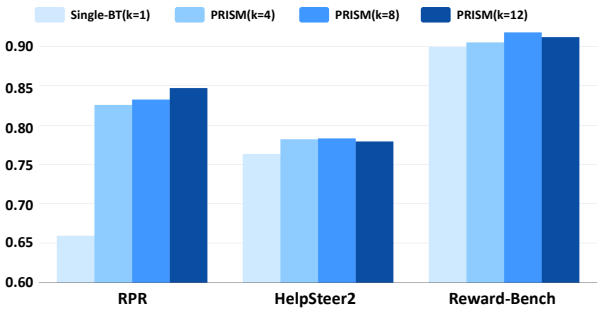

Figure 8: Capability profiles of different K values across three datasets.

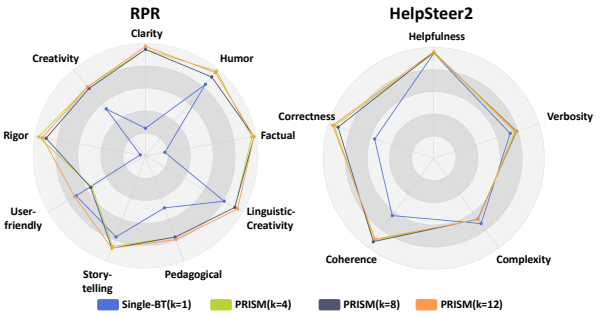

Figure 9: Radar chart comparison of different heads across attribute dimensions on RPR and HelpSteer2.

C.4.3 Expert Ability Explanation

After training the model, we can assign a certain meaning to each head afterwards, as shown in the

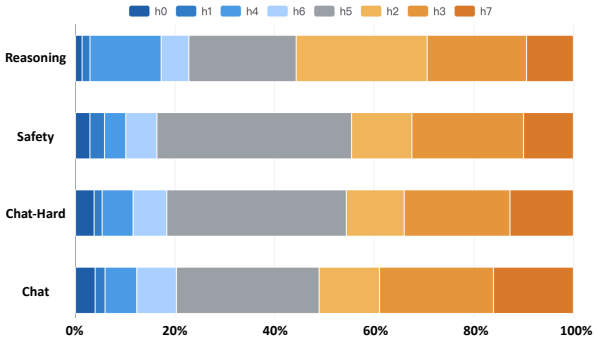

Figure 10: Distribution of routing across attribute subsets on Reward-Bench. The orange head group represents objective-oriented expert heads, while the blue head group represents subjective-oriented expert heads.

Table 13. These experts can be divided into two groups: subjective experts and objective experts.

Table 11: Accuracy scores on **RPR** test set. On average, PRISM outperforms baselines across various attributes and overall results. All baselines use the same 8B base model.

Method	Clarity	Creativity	Scientific Rigor	User-Friendliness	Storytelling	Pedagogical	Linguistic Creativity	Factual Accuracy	Humor	Average
Single Head	0.5094	0.5972	0.3333	0.7865	0.8125	0.5484	0.7596	0.3380	0.9167	0.6224
Static Mixture	0.3962	0.5641	0.3214	0.7753	0.7375	0.6129	0.8462	0.3380	0.8571	0.6055
ARMO	0.9057	0.6806	0.9405	0.6966	0.7875	0.7903	0.9135	0.9014	0.9463	0.8403
MiCRo	0.8189	0.7750	0.8500	0.8225	0.8725	0.8065	0.8654	0.8620	0.9214	0.8438
PRISM(Sky-3B)	0.9245	0.6111	0.9405	0.7303	0.7500	0.7097	0.8654	0.9155	0.8929	0.8169
PRISM	0.9434	0.6667	0.9286	0.7528	0.7625	0.8710	0.8558	0.9859	0.9167	**0.8498**

Table 12: Accuracy scores on **HelpSteer3** test set. On average, PRISM outperforms baselines across various attributes and overall results. All baselines use the same 8B base model.

Method	Helpfulness	Correctness	Coherence	Complexity	Verbosity	Average
Single Head	0.7636	0.7318	0.6909	0.7682	0.7818	0.7473
Static Mixture	0.7818	0.7364	0.7136	0.7455	0.7636	0.7482
ARMO	0.6919	0.6395	0.7593	0.7132	0.7500	0.7108
MiCRo	0.7864	0.7227	0.7242	0.7727	0.7712	0.7555
PRISM(Sky-3B)	0.7343	0.7365	0.7241	0.7434	0.8355	0.7545
PRISM	0.7766	0.7682	0.7802	0.7325	0.8289	**0.7783**

Table 13: Head-level attribute specialties, abstracted information patterns, and the resulting grouping structure.

Head	**Head Specialties (Attributes)**	**Abstracted Information Pattern**	**Group**
0	Complexity, Verbosity	Redundancy & conciseness (compression and redundancy control).	Subjective
1	Complexity, Verbosity	Redundancy & conciseness (compression and redundancy control).	Subjective
2	Helpfulness, Correctness, Factual	Accuracy & utility (factually correct and practically helpful content).	Objective
3	Helpfulness, Coherence, Clarity, User-friendly, Factual	Clarity & structuring (well-organized, coherent, and easy-to-follow responses).	Objective
4	Creativity, StoryTelling, Linguistic-Creativity	Creativity & affect (creative expression and narrative/linguistic novelty).	Subjective
5	Correctness, Clarity, Rigor, User-friendly, Pedagogical-effectiveness, Factual	Rigor & pedagogy (rigorous, instructional, and user-comprehensible explanations).	Consensus
6	Creativity, StoryTelling, Linguistic-Creativity, Humor	Humor & entertainment (humorous, engaging, and creative storytelling).	Subjective
7	Coherence, Clarity, Rigor, Factual	Rigor & factuality (logically tight and fact-consistent reasoning).	Objective