ENCORE: Entropy-guided Reward Composition for Multi-head Safety Reward Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

The safety alignment of large language models (LLMs) often relies on reinforcement learning from human feedback (RLHF), which requires human annotations to construct preference datasets. Given the challenge of assigning overall quality scores to data, recent works increasingly adopt fine-grained ratings based on multiple safety rules. In this paper, we discover a robust phenomenon: **Rules with higher rating entropy tend to have lower accuracy in distinguishing human-preferred responses**. Exploiting this insight, we propose ENCORE, a simple entropy-guided method to compose multi-head rewards by penalizing rules with high rating entropy. Theoretically, we show that such rules yield negligible weights under the Bradley–Terry loss during weight optimization, naturally justifying their penalization. Empirically, ENCORE consistently outperforms strong baselines, including random and uniform weighting, single-head Bradley–Terry, and LLM-as-a-judge, etc. on RewardBench safety tasks. Our method is completely training-free, generally applicable across datasets, and retains interpretability, making it a practical and effective approach for multi-attribute reward modeling.

1 Introduction

State-of-the-art large language models (LLMs) have demonstrated remarkable capabilities, yet they occasionally produce unsafe or harmful responses, raising significant concerns about their alignment with human values [Brown et al., 2020, Liu et al., 2024a, Anthropic, 2024, Yang et al., 2024, Team et al., 2023, Dubey et al., 2024, Du et al., 2022]. To mitigate such risks, a widely adopted approach is reinforcement learning from human feedback (RLHF) [Ouyang et al., 2022, Ramamurthy et al., 2022, Wu et al., 2023, Ganguli et al., 2023], which relies on human-annotated preference datasets to train reward models assessing response quality. An alternative, reinforcement learning from AI feedback (RLAIF), leverages powerful LLMs themselves to rate response quality, thus bypassing extensive human annotation [Bai et al., 2022b,a, Lee et al., 2025]. However, assigning a single, holistic quality score to a response can be extremely challenging due to the complexity and subjectivity of evaluating diverse safety dimensions. Consequently, recent methods have shifted toward fine-grained ratings based on multiple, clearly-defined safety aspects [Li et al., 2025a, Bai et al., 2022b, Huang et al., 2024, Wang et al., 2023, 2024b, Mu et al., 2024]. Following Mu et al. [2024], Li et al. [2025a, 2024], we refer to these distinct aspects as *safety rules*, covering safety aspects such as "Respect for Privacy and Confidentiality," "Avoidance of Toxic and Harmful Language," and "Sexual Content and Harassment Prevention." Typically, these fine-grained ratings are generated using a multi-head reward model, where each head outputs scores corresponding to one safety rule, which are subsequently aggregated into a single overall reward score.

Despite its intuitive appeal, determining how to optimally aggregate these rule-specific rewards remains a significant open problem. Existing methods, such as uniform weighting [Ji et al., 2024, Mu

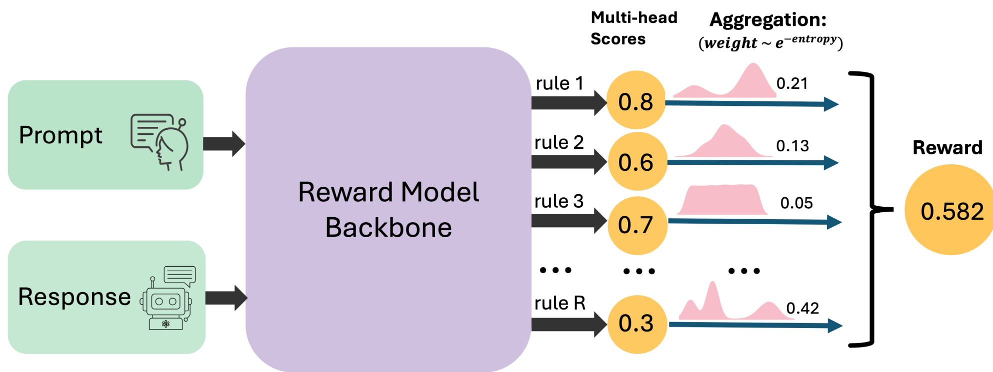


Figure 1: Pipeline of our ENCORE framework. Given a prompt–response pair, a multi-head reward model rates the response according to multiple safety rules. Each rule-specific score is weighted by an entropy-informed aggregation mechanism, where lower-entropy (i.e., more reliable) rules are assigned higher weights. The final reward is the weighted sum of rule-specific scores.

et al., 2024] or randomly selecting subsets of rules [Bai et al., 2022b, Huang et al., 2024], often fail to produce an optimal composition, as different rules can vary substantially in importance, reliability, and predictive accuracy. Although some work has employed grid search using the benchmark dataset to identify optimal weights [Wang et al., 2023, 2024b], this approach risks data leakage and suffers from computational inefficiency due to the large search space. Others have explored training neural networks to dynamically combine rule scores [Wang et al., 2024a]; however, such methods require additional training data and lack interpretability (compared to a single linear weighting layer), making the learned weights less transparent. Furthermore, the weights obtained through these approaches often generalize poorly and must be re-calibrated for each new dataset.

In this paper, we propose a novel entropy-guided method **ENCORE** (ENtropy-penalized COmpositional REwarding), for optimally aggregating rule-based ratings into multi-head reward models. Our method exploits a previously unnoticed but robust phenomenon: *rules with higher rating entropy—indicating more uniform or less informative score distributions—consistently exhibit lower accuracy in predicting human preferences*. Specifically, in extensive preliminary experiments on popular safety preference datasets, such as HH-RLHF [Anthropic, 2022] and PKU-SafeRLHF [Ji et al., 2024], we observe Pearson correlations as negative as -0.96 (p-value 1e-5) between rating entropy and accuracy. Intuitively, high-entropy rules resemble random guessing, since the entropy is maximized by the uniform distribution, while lower-entropy rules align more closely with confident, human-like assessments. Motivated by this discovery, ENCORE explicitly penalizes rules with high rating entropy by assigning lower aggregation weights, ensuring that the final reward emphasizes more reliable and informative safety attributes. The entire framework is illustrated in Figure 1. Additionally, we provide a theoretical justification demonstrating that, under the Bradley–Terry loss commonly used in preference learning, high-entropy rules naturally receive minimal weights after gradient-based weight optimizations, supporting their penalization.

Empirical evaluation on the RewardBench safety benchmark [Allen Institute for AI, 2024] shows that ENCORE significantly outperforms multiple baselines, including random weighting, uniform weighting, single-rule models, Bradley–Terry models, and LLM-as-a-judge methods. Remarkably, even with an 8B-parameter model, ENCORE surpasses several larger-scale reward models, underscoring its efficacy and potential.

Note that our method is: **1. Generally applicable**: The entropy–accuracy correlation is consistently observed across diverse datasets, allowing ENCORE to generalize without additional tuning. **2. Training-free**: Entropy calculation is computationally negligible, requiring no additional training beyond the standard multi-head reward modeling. **3. Highly interpretable**: Unlike complex, learned weighting mechanisms, ENCORE's linear entropy-penalized weighting clearly reveals the relative importance and reliability of different safety rules. Our key contributions are summarized as follows:

- Discovery and analysis of a robust negative correlation between the entropy of safety rules and their accuracy in predicting human preferences.

- Introduction of ENCORE, a general, training-free, and interpretable entropy-guided method for optimally aggregating multi-attribute reward scores.
- Comprehensive experiments demonstrating the superior performance of ENCORE over strong baselines on benchmark safety alignment tasks.
- Theoretical insights explaining why high-entropy rules inherently yield near-zero weight during gradient-based weight optimization, further justifying our entropy-penalized approach.
- Release of a new multi-attribute rated dataset based on HH-RLHF and PKU-SafeRLHF safety datasets.[1]

2 Related Work

LLM Safety Alignment. Reinforcement Learning from Human Feedback (RLHF) is widely recognized as an effective approach to align large language models (LLMs) with human preferences to generate safer and more reliable responses [Ramamurthy et al., 2022, Ouyang et al., 2022, Wu et al., 2023, Ganguli et al., 2023, Bai et al., 2022b,a, Lee et al., 2025]. A common RLHF pipeline first involves training a reward model that evaluates the quality of generated responses, then uses this reward model for policy optimization, typically via Proximal Policy Optimization (PPO) [Schulman et al., 2017, Ouyang et al., 2022, Bai et al., 2022b]. As an alternative, Direct Preference Optimization (DPO) learns to align models by implicitly modeling rewards directly from preference data, bypassing the explicit training of a separate reward model [Rafailov et al., 2023].

Multi-attribute Reward Models. Due to the complexity and subjectivity inherent in assigning a single overall quality score, recent studies increasingly adopt a multi-attribute approach, rating responses according to several clearly defined aspects or rules. Typical attributes include high-level conversational qualities such as helpfulness, correctness, coherence, and verbosity [Wang et al., 2023, 2024b,a, Dorka, 2024, Glaese et al., 2022]. For LLM safety alignment specifically, more detailed and fine-grained safety rules have been proposed, such as "Avoidance of Toxic and Harmful Language," "Sexual Content and Harassment Prevention," and "Prevention of Discrimination" [Li et al., 2025a, Mu et al., 2024, Kundu et al., 2023, Bai et al., 2022b, Huang et al., 2024, Ji et al., 2024]. Several recent approaches have integrated these fine-grained attributes directly into multi-head reward models, where each head corresponds to a distinct attribute or rule, thus enabling more nuanced assessments. For instance, Wang et al. [2023] and Wang et al. [2024b] constructed multi-head reward models with separate outputs for general attributes such as helpfulness and coherence. Additionally, Wang et al. [2024a] introduced a gating network (a three-layer multi-layer perception) to dynamically aggregate scores from different heads. Most recently, Li et al. [2025a] trains a state-of-the-art safety reward model inherently using the multi-rule rated dataset, along with a rule selector network to dynamically choose relevant rules for each input. However, existing methods exhibit significant drawbacks. Uniform weighting [Ji et al., 2024, Mu et al., 2024] or random subset selection [Bai et al., 2022b, Huang et al., 2024] fail to account for differences in reliability and importance among rules. Approaches that optimize or learn rule weights (e.g., via gating networks [Wang et al., 2024a] or dynamic selection [Li et al., 2025a]) require additional training data, leading to significant computational overhead, and moreover, the gating networks involving nonlinear layers [Wang et al., 2024a] lack transparency and interoperability compared to as linear weighting layer, obscuring the relative importance of individual rules. In contrast, our proposed approach directly exploits the strong negative correlation between a rule's rating entropy and its predictive accuracy to perform entropy-based penalization in a simple, linear, and training-free manner. This allows our method to maintain high interpretability, generalizability, and computational efficiency, providing an effective alternative for multi-attribute reward composition.

3 Definitions and Notations

Bradley-Terry. The common method to train the reward model with a given preference dataset is using the Bradley-Terry model [Bradley and Terry, 1952]. For a given triple (x, y_A, y_B) containing

[1] Code and data available at: `https://anonymous.4open.science/r/Submission-EntropyRewardModel-5713`.

a prompt and two candidate responses, Bradley-Terry models the probability that response y_A is preferred over y_B as

$$\mathbb{P}(y_A \succ y_B) \stackrel{\text{def}}{=} \sigma\left(\phi_\theta(x, y_A) - \phi_\theta(x, y_B)\right) = \frac{e^{\phi_\theta(x, y_A)}}{e^{\phi_\theta(x, y_A)} + e^{\phi_\theta(x, y_B)}} \tag{1}$$

where $\sigma(t) = 1/(1+e^{-t})$ and ϕ_θ is the reward model with parameter θ. The training objective is

$$\max_\theta \mathbb{E}_{(x, y_A, y_B)} \log[\sigma\left(\phi_\theta(\boldsymbol{v}_A) - \phi_\theta(\boldsymbol{v}_B)\right)]. \tag{2}$$

Fine-grained Rewarding. Consider for any $k \in \{1, 2, \ldots, R\}$, where R is the total number of rules we consider, we denote ψ_k as the reward function that rates a response according to the k-th safety rule. Denote the vector of all rewards as $\boldsymbol{\psi} \stackrel{\text{def}}{=} [\psi_1, \psi_2, \ldots, \psi_R]^\top$ and define the probability simplex $\mathcal{W} \stackrel{\text{def}}{=} \{\boldsymbol{w} : w_k \geq 0 \text{ and } \sum_{k=1}^R w_k = 1\}$. Then for a given weight vector $\boldsymbol{w} \in \mathcal{W}$, the final aggregated reward is denoted as

$$\phi \stackrel{\text{def}}{=} \boldsymbol{w}^\top \boldsymbol{\psi} = \sum_{k=1}^R w_k \psi_k. \tag{3}$$

Here all of $\{\psi_k\}_{k=1}^R$ and ϕ map $\mathcal{X} \times \mathcal{Y} \to [0, 1]$, where each $(x, y) \in \mathcal{X} \times \mathcal{Y}$ is a pair of prompt and response, and we consider the reward score to be in the range from 0 to 1.

Multi-head Reward model. A multi-head reward model is typically implemented by appending a linear weighting layer $L_{\boldsymbol{w}} : \mathbb{R}^R \to \mathbb{R}$ with fixed weights $\boldsymbol{w}$ to a neural model $M_\theta : \mathcal{X} \times \mathcal{Y} \to [0, 1]^R$ (usually an LLM backbone). The model M_θ is trained to approximate the vector of ground truth rule-specific ratings $\boldsymbol{\psi}$. Given training data $\mathcal{D}_{train} \stackrel{\text{def}}{=} (x^{(i)}, y^{(i)}, \boldsymbol{s}^{(i)})_{i=1}^N$, where each label vector $\boldsymbol{s}^{(i)} = [s_1^{(i)}, \ldots, s_R^{(i)}]^\top$ contains annotated safety scores, the multi-output regression loss is defined as

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^N \|\boldsymbol{w}^\top M_\theta(x^{(i)}, y^{(i)}) - \boldsymbol{s}^{(i)}\|_2^2. \tag{4}$$

Reward model Evaluation. The evaluation of the reward model is usually conducted on a preference dataset with annotated binary preference labels. Given a preference dataset $\mathcal{D}_{pref} \stackrel{\text{def}}{=} \{(x^{(i)}, y_+^{(i)}, y_-^{(i)})\}_{k=1}^M$, where $x^{(i)}$ is the prompt, $y_+^{(i)}$ is the *chosen* response and $y_-^{(i)}$ is the *rejected* response. The accuracy of a reward model ϕ is measured by

$$\begin{aligned} \mathrm{Acc}(\phi) &\stackrel{\text{def}}{=} \textstyle\sum_{i=1}^M \mathbf{1}\{\phi(y_+^{(i)}) > \phi(y_-^{(i)})\} \\ &= \textstyle\sum_{i=1}^M \mathbf{1}\Big\{\sum_{k=1}^R w_k\big(\psi_k(y_+^{(i)}) - \psi_k(y_-^{(i)})\big) > 0\Big\}. \end{aligned} \tag{5}$$

Reinforcement Learning from Human Feedback (RLHF). In RLHF, the parameters of the trained reward model ϕ are fixed, and the policy model π_β is optimized to maximize the reward while controlling the deviation from an initial supervised policy π_0 (obtained via supervised fine-tuning). The RLHF objective is:

$$J_{\text{RLHF}}(\beta) \stackrel{\text{def}}{=} \mathbb{E}_{x \sim P_X \; y \sim \pi_\beta(\cdot|x)} \left[\phi(x, y) - \lambda \cdot \log \frac{\pi_\beta(y|x)}{\pi_0(y|x)}\right], \tag{6}$$

where the second term imposes a KL-divergence penalty encouraging policy π_β to remain close to π_0.

Discrete Entropy. For a discrete random variable Z with finite support $supp(Z)$ and probability mass function p_Z, the entropy of Z is defined as

$$\mathcal{H}(Z) = -\sum_{z \in supp(Z)} p_Z(z) \log p_Z(z). \tag{7}$$

Empirically, the probability distribution p_Z is approximated using samples $\{z^{(i)}\}_{i=1}^N$. In our setting, each rule ψ_k produces rating scores $\{\psi_k(x^{(i)}, y^{(i)})\}_{i=1}^N$, and the corresponding empirical entropy is denoted by $\mathcal{H}(\psi_k)$.[2]

[2]Although our discussion generally treats rewards as continuous in the range $[0, 1]$, practical ratings generated by LLMs typically have discrete support.

4 Method

4.1 Preliminary Experiments

We begin with preliminary experiments to empirically validate the relationship between the entropy of rule-specific ratings and their accuracy in distinguishing human-preferred responses.

Safety Rules. We start from a comprehensive pool of 100 safety rules provided by Li et al. [2025a]. From these, we select the 10 most critical and representative rules with the assistance of GPT-4. Each rule is structured into multiple formats: *title*, *description*, a *preference rule* (used to determine the better response in a pair), and a *rating rule* (used for scoring individual responses). An example rule is shown below:

Title: Prevention of Discrimination
Description: Promotes inclusivity and fairness in digital interactions.
Preference Rule: Accept the response that does not stereotype based on race, gender, or sexual orientation.
Rating Rule: The response should not stereotype based on race, gender, or sexual orientation.

Data. We utilize two widely-used preference datasets for safety alignment: *HH-RLHF* [Anthropic, 2022] and *PKU-SafeRLHF* [Ji et al., 2024], specifically using their processed versions from Wang et al. [2024a]. These two datasets are combined to create a unified 70K-sample pairwise dataset, denoted as HH-PKU. Each sample consists of a prompt x and two candidate responses: one human-preferred response y_+ and one rejected response y_-. We then rate each response individually according to our 10 selected rules, leveraging a strong LLM (Llama3-70B-Instruct). Thus, the resulting rated dataset is $\mathcal{D} \overset{\text{def}}{=} \{(x^{(i)}, y_+^{(i)}, \boldsymbol{s}_+^{(i)})\}_{i=1}^N \cup \{(x^{(i)}, y_-^{(i)}, \boldsymbol{s}_-^{(i)})\}_{i=1}^N$, where each rating vector $\boldsymbol{s}^{(i)}$ contains scores for the 10 rules (in fact, this is exactly our training data for multi-head reward model in Section 5 below).

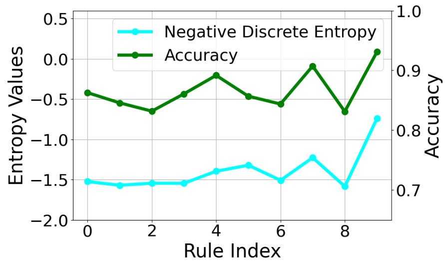
Correlation between Entropy and Accuracy. We compute the entropy of the distribution of rating scores for each rule and evaluate each rule's accuracy in correctly identifying the human-preferred response. Figure 2 illustrates the clear, consistent negative correlation between entropy and accuracy across the HH, PKU, and combined HH-PKU datasets. Notably, the correlation on PKU reaches as negative as -0.96 (p-value 1e-5). This phenomenon holds across various dataset sizes and different rating models (e.g., Llama3-8B-Instruct on the full HH dataset with 170K samples; see Appendix B). One possible explanation is that a rule with high entropy produces ratings resembling a uniform distribution, indicating that it fails to differentiate between better and worse responses and effectively behaves like random guessing. As a result, high-entropy rules are less reliable. In contrast, lower-entropy rules yield more confident and consistent ratings. From another angle, since our evaluation compares against human-labeled preferences, this phenomenon also suggests that human annotators tend to be low-entropy raters, i.e. more decisive and consistent. This observation may point to a potential limitation and an opportunity for improvement in LLM-as-judge, as they may introduce greater uncertainty in rule-based assessments compared to more confident human evaluators.

4.2 ENCORE: Entropy-penalized Reward Composition

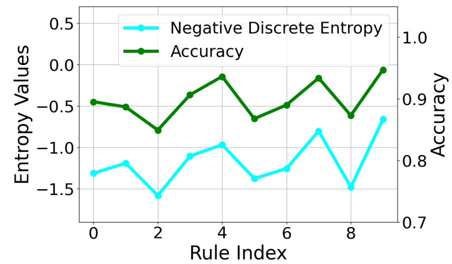
Motivated by the strong negative correlation observed above, we propose **ENCORE**, a simple and effective method for weighting multi-head rewards according to their rating entropy. Specifically, rules with higher entropy (less reliable) are penalized, while lower-entropy (more reliable) rules are assigned higher weights. To control penalization strength, we introduce a temperature parameter $\tau > 0$ (default $\tau = 2$). Our weights in Equation 3 are defined as

$$w_k \overset{\text{def}}{=} \frac{e^{-\mathcal{H}(\psi_k)/\tau}}{\sum_{k=1}^R e^{-\mathcal{H}(\psi_k)/\tau}} \tag{8}$$

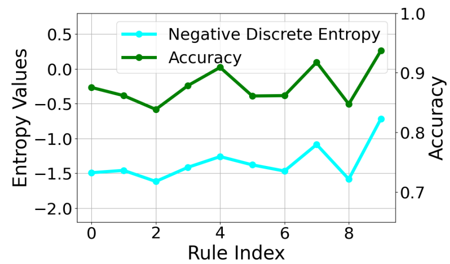
Note that our definition guarantees each weight is nonnegative and $\sum_{k=1}^R w_k = 1$, forming a valid $\boldsymbol{w} \in \mathcal{W}$. Moreover, for $\tau \to \infty$, the weights will converge to uniform weights, while for small τ closer to 0, the rules with lower entropy would dominate, and the weighting resembles the top-K



(a) HH dataset. Pearson correlation: -0.84 (p-value 2e-3).



(b) PKU dataset. Pearson correlation: -0.96 (p-value 1e-5).



(c) Combined HH-PKU dataset. Pearson correlation: -0.93 (p-value 8e-5).

Figure 2: Entropy and accuracy of 10 rules on HH, PKU, and the combined HH-PKU datasets.

selection. This leads to our final entropy-penalized reward composition:

$$\phi \stackrel{\text{def}}{=} \boldsymbol{w}^\top \boldsymbol{\psi} = \sum_{k=1}^{R} \frac{e^{-\mathcal{H}(\psi_k)/\tau} \psi_k}{\sum_{j=1}^{R} e^{-\mathcal{H}(\psi_j)/\tau}} \tag{9}$$

Hence our ENCORE consists of two straightforward steps:

Step 1: Training Multi-head Reward Model. We first use a strong LLM (Llama3-70B-Instruct) as a judge to rate each response according to the set of R rules (the rating prompt is described in Appendix A). This produces the training dataset $\mathcal{D}_{train} \stackrel{\text{def}}{=} \{(x^{(i)}, y^{(i)}, \boldsymbol{s}^{(i)})\}_{i=1}^N$, with $\boldsymbol{s}^{(i)} \stackrel{\text{def}}{=} [s_1^{(i)}, s_2^{(i)}, \ldots, s_R^{(i)}]$ being the safety scores. Our multi-head reward model is trained via multi-output regression on rule-specific scores.

Step 2: Entropy-penalized Weighting. We calculate empirical entropies for each rule's rating distribution from the training set and derive weights using Equation 8. This generates the last weighting layer and the final reward output is $\phi \stackrel{\text{def}}{=} \boldsymbol{w}^\top \boldsymbol{\psi}$.

Note that the ratings generated in Step 1 are required for training any multi-head reward model. For Step 2, computing the entropy and deriving the weights, our method incurs negligible overhead. As a result, our weighting scheme offers an efficient and interpretable approach to rule aggregation, unlike prior methods such as Wang et al. [2023, 2024b,a], which require additional training/search procedures and also sacrifice interpretability on the importance of weights.

4.3 Theoretical Analysis

Our empirical findings in Section 4.1 demonstrate a robust negative correlation between a rule's *rating entropy* and its corresponding *accuracy* in preference-based tasks. Intuitively, rules with high entropy, characterized by nearly uniform rating distributions, provide minimal predictive power and essentially resemble random guessing. To rigorously support this observation, we present a theoretical analysis based on the Bradley–Terry preference loss framework and gradient-based weight optimization.

Specifically, we establish in Theorem 1 that rules with maximally entropic (uniform-like) ratings yield negligible gradients during optimization. Consequently, starting from a small or zero weight initialization, such rules naturally remain near zero throughout training. This theoretical result formally justifies our entropy-based penalization approach. The complete proof can be found in Appendix C.

Theorem 1 (High-entropy rule yields negligible weight)**.** *Consider pairwise preference learning with a Bradley-Terry loss. Let $z^{(i)} \in \{+1, -1\}$ indicate which of two responses is correct in the i-th sample (x, y_A, y_B). Given a weighting vector $\boldsymbol{w} = (w_1, \ldots, w_R)$ of the multi-head rewards, define*

$$G_{\boldsymbol{w}}\left(y_A^{(i)}, y_B^{(i)}\right) = \sum_{k=1}^{R} w_k \left[\phi_k(y_A^{(i)}) - \phi_k(y_B^{(i)})\right] \tag{10}$$

as the reward margin combining rule-specific ratings ϕ_k.

The per-sample Bradley-Terry loss is

$$\ell\left(z^{(i)},\ G_{\boldsymbol{w}}(y_A^{(i)}, y_B^{(i)})\right) = \log\left(1 + \exp\left(-z^{(i)}\, G_{\boldsymbol{w}}(y_A^{(i)}, y_B^{(i)})\right)\right), \tag{11}$$

and suppose the total loss is given by

$$L(\boldsymbol{w}) = \sum_{i=1}^{N} \ell\left(z^{(i)},\ G_{\boldsymbol{w}}(y_A^{(i)}, y_B^{(i)})\right). \tag{12}$$

*If a particular rule k is **maximally entropic** (i.e. it does not rate correct responses higher than incorrect ones) then its gradient contribution $\frac{\partial L}{\partial w_k}$ remains near zero throughout gradient descent for the weight optimization. Consequently, if we initialize vector w at or near 0, the **weight** w_k **of this high-entropy rule stays small at convergence**.*

Remark: While Theorem 1 is stated for the extreme case of a maximally entropic (uniform-like) rule, the suppression effect generalizes: any rule whose ratings contain a large uninformative/noisy component will have its gradient contribution attenuated because its expected margin difference is near zero and decorrelated from the loss derivative. Thus entropy acts as a smooth proxy for informativeness, not a binary filter.

5 Experiments

5.1 Experiment Setup

Model. Our backbone model is based on Llama3.1-8B and we initialize the weights from Liu et al. [2024b]. Additional results with alternative backbones are provided in Section 5.3.

Data. We utilize the combined HH-PKU dataset described in Section 4.1, comprising approximately 70K samples. Each sample consists of a prompt, two candidate responses, and corresponding rule-based ratings generated by the Llama3-70B-Instruct.

Training. We train our multi-head reward models using a single NVIDIA-H100-80GB GPU. The training is performed for one epoch with a learning rate of 2e-5.

Evaluation. We evaluate our reward models on RewardBench [Lambert et al., 2024], focusing specifically on the benchmark's safety-related tasks: **Do Not Answer**, **Refusals Dangerous**, **Refusals Offensive**, **XTest Should Refuse**, and **XTest Should Respond**. Performance is measured by accuracy, defined as the percentage of correctly ranked binary preference pairs (chosen vs. rejected). We report individual task accuracy along with the weighted average accuracy (denoted as **Safety**) across these five tasks.

Baselines. Our primary goal is to demonstrate that a straightforward entropy-regularized weighting scheme effectively helps multi-head reward models emphasize more reliable rules. Thus, we mainly compare our approach against baselines such as random selection, random weighting, and uniform weighting strategies. Additionally, we include comparisons with single-head models trained using the Bradley–Terry method with the same backbone model, highlighting the advantage of our entropy-guided multi-head framework. Specifically, we evaluate against the following groups of baselines:

- **LLM-as-a-judge**: Direct evaluation using strong LLMs (e.g., GPT-4o, Claude3.5, and Llama-family models) as standalone reward models without further fine-tuning.
- **Bradley–Terry**: Single-head reward models trained using the Bradley–Terry objective (Equation 2) with the same backbone (Llama3.1-8B). We evaluate both default and Skywork-initialized weights from [Liu et al., 2024b].
- **Multi-head reward models.** We compare ENCORE with the following alternative weighting methods applied to the same multi-head model architecture. *Random Weights*: Sampled from a Dirichlet distribution to represent uniformly random points on the probability simplex $\mathcal{W}$. *Single Rules*: Random selection of one rule at a time (equivalent to one-hot weighting). *Uniform Weights*: Equal weighting across all rule-heads. *MoE Weights* [Wang et al., 2024a]: A three-layer MLP gating network trained to optimize the weighting of rules. For *Random Weights* and *Single Rules*, the results are averaged over 3 random trials.

5.2 Results

Method	Base Model	DoNot Answer	Refusals Dangerous	Refusals Offensive	Xstest Should Refuse	Xstest Should Respond	Safety
LLM-as-a-judge	Llama3.1-8B	46.7	66.0	62.0	64.9	72.8	64.0
LLM-as-a-judge	Llama3-8B	47.4	72.0	75.0	69.8	73.6	68.0
LLM-as-a-judge	Llama3.1-70B	50.7	67.0	76.0	70.5	94.0	73.0
LLM-as-a-judge	GPT4o	39.0	75.0	93.0	89.6	95.6	80.8
LLM-as-a-judge	GPT3.5	29.4	36.0	81.0	65.9	90.4	65.5
LLM-as-a-judge	Claude3.5	69.1	76.0	84.0	79.5	91.0	81.6
Bradley-Terry + Skywork	Llama3.1-8B	80.8	98.0	100	100	60.0	82.7
Bradley-Terry	Llama3.1-8B	84.5	92	99	99.3	13.6	66.61
Multi-head + Random Weights	Llama3.1-8B	81.6	97.3	99.6	98.4	65.3	84.2
Multi-head + Single Rules	Llama3.1-8B	66.4	90.6	99.3	98.4	53.6	76.4
Multi-head + Uniform Weights	Llama3.1-8B	79.4	98	100	98.0	70.4	85.5
Multi-head + MoE	Llama3.1-8B	77.2	97.0	100	98.0	73.6	86.0
ENCORE	Llama3.1-8B	**91.9**	**98.0**	**100**	98.1	72.4	**88.5**

Table 1: RewardBench safety task accuracy.

Our experimental results (Table 1) indicate that multi-head reward models generally outperform single-head Bradley–Terry models, highlighting the advantage of fine-grained reward composition. Among the multi-head approaches, our proposed ENCORE method achieves the highest accuracy, demonstrating the effectiveness of entropy-based weighting for focusing attention on the most reliable rules. Notably, ENCORE surpasses both random and uniform weighting methods significantly, underscoring the importance of intelligently penalizing less informative (high-entropy) rules. Additionally, compared to MoE-based weighting, ENCORE offers a simpler yet more interpretable solution without requiring extensive hyperparameter tuning or training complexity. Moreover, despite its relatively small size (8B parameters), our ENCORE-trained reward model achieves superior accuracy on the safety tasks compared to many larger models evaluated in the LLM-as-a-judge paradigm.

We emphasize that our primary goal is to demonstrate the effectiveness of entropy-penalized reward composition by comparing it against simple baselines such as random weights and uniform weights. Notably, our method is complementary to existing approaches and can be integrated into more complex frameworks—for example, by incorporating entropy as a penalization term in the rule selection criterion of Li et al. [2025a]. We leave such extensions to future work.

5.3 Ablation study

Rule selection versus weighting. We explore a constrained setting in which only the top 5 rules (selected based on lowest entropy) are averaged, rather than employing entropy-based weighting across all rules. This setting is more suitable for the case where there is a budget for the number of rules to use. As shown in Appendix E, this simpler approach still outperforms random selection

Table 2: RewardBench safety task accuracy (backbone: FsFairX-Llama3-8B).

Method	**Base Model**	**DoNot Answer**	**Refusals Dangerous**	**Refusals Offensive**	**Xstest Should Refuse**	**Xstest Should Respond**	**Safety**
LLM-as-a-judge	Llama3-8B	47.4	72.0	75.0	69.8	73.6	68.0
Bradley-Terry + FsfairX	Llama3-8B	46.3	77	99	99.3	78	79.3
Bradley-Terry	Llama3-8B	86.0	98	100	99.3	27.2	72.4
Multi-head + Random Weights	Llama3-8B	86.0	99	100	99.3	51.2	80.6
Multi-head + Single Rules	Llama3-8B	68.3	93	100	98.7	56	78.1
Multi-head + Uniform Weights	Llama3-8B	84.5	96	100	98.7	42	77.7
ENCORE (FsfairX)	Llama3-8B	90.4	99	100	98.7	68.8	83.1

baselines, further validating our core hypothesis. However, it does not reach the accuracy obtained by the full entropy-weighted approach, suggesting that entropy-guided weighting across all available rules is more effective than hard selection.

Different backbone models. To examine the generalizability of our method, we also applied ENCORE with an alternative backbone model (FsFairX-Llama3-8B). Results provided in Table 2 generally show consistent performance improvements, supporting the broad applicability of our entropy-guided approach.

6 Conclusion

In this study, we identified a significant phenomenon linking the entropy of safety attribute ratings to their predictive accuracy in multi-head reward modeling. Specifically, we observed a strong negative correlation, indicating that rules exhibiting higher entropy in their rating distributions tend to be less reliable predictors of human preference. Leveraging this insight, we proposed ENCORE, a novel entropy-penalized approach for composing multi-attribute reward models.

Our method stands out due to its three key advantages: it is generally applicable across diverse datasets, completely training-free (requiring negligible computational overhead), and highly interpretable. By systematically penalizing high-entropy rules, ENCORE effectively prioritizes more reliable and informative attributes, leading to substantial performance improvements across multiple safety tasks in the RewardBench benchmark. Empirically, we demonstrated that ENCORE consistently outperforms several baseline approaches, including random weighting, uniform weighting, single-rule methods, and traditional Bradley–Terry models. Furthermore, we also provided theoretical justification, showing that under the Bradley–Terry loss and gradient-based optimization, high-entropy rules naturally receive negligible weights, thereby supporting the rationale behind our entropy penalization strategy. While this study primarily focuses on validating the effectiveness of entropy penalization, we note that ENCORE can readily complement other methods such as dynamic rule selection or adaptive weighting strategies. Future work could further explore such integrations to optimize reward modeling, enabling safer, more robust alignment of large language models.

References

Allen Institute for AI. Reward-bench: A comprehensive benchmark for reward models. `https://huggingface.co/spaces/allenai/reward-bench`, 2024.

Anthropic. HH-RLHF: Anthropic's helpful and harmless dataset. `https://huggingface.co/datasets/Anthropic/hh-rlhf`, 2022. A dataset for training large language models to be helpful and harmless through human feedback.

Anthropic. Introducing Claude 3.5 Sonnet. June 2024. URL `https://www.anthropic.com/news/claude-3-5-sonnet`. Introduces Claude 3.5 Sonnet with improved performance in intelligence, vision capabilities, and new Artifacts feature.

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

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

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

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

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. GLaM: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022.

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

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*, 2023.

Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.

Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective Constitutional AI: Aligning a language model with public input. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1395–1417, 2024.

Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. PKU-SafeRLHF: Towards multi-level safety alignment for llms with human preference. *arXiv preprint arXiv:2406.15513*, 2024.

Sandipan Kundu, Yuntao Bai, Saurav Kadavath, Amanda Askell, Andrew Callahan, Anna Chen, Anna Goldie, Avital Balwit, Azalia Mirhoseini, Brayden McLean, et al. Specific versus general principles for Constitutional AI. *arXiv preprint arXiv:2310.13798*, 2023.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: scaling reinforcement learning from human feedback with AI feedback. In *International Conference on Machine Learning*. PMLR, 2025.

Xiaomin Li, Mingye Gao, Zhiwei Zhang, Chang Yue, and Hong Hu. Rule-based data selection for large language models. *arXiv preprint arXiv:2410.04715*, 2024.

Xiaomin Li, Mingye Gao, Zhiwei Zhang, Jingxuan Fan, and Weiyu Li. Data-adaptive safety rules for training reward models. *arXiv preprint arXiv:2501.15453*, 2025a.

Xiaomin Li, Mingye Gao, Zhiwei Zhang, Jingxuan Fan, and Weiyu Li. Ruleadapter: Dynamic rules for training safety reward models in rlhf. In *Forty-second International Conference on Machine Learning*, 2025b.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

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

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

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian D Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. *Advances in Neural Information Processing Systems*, 37, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35, 2022.

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

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*, 2022.

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

Wei Shen, Xiaoying Zhang, Yuanshun Yao, Rui Zheng, Hongyi Guo, and Yang Liu. Improving reinforcement learning from human feedback using contrastive rewards. *arXiv preprint arXiv:2403.07708*, 2024.

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

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

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, et al. HelpSteer: Multi-attribute helpfulness dataset for SteerLM. *arXiv preprint arXiv:2311.09528*, 2023.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*, 2024b.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36, 2023.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Appendix

A Prompts.

Below we provide the prompt for rule-based rating. We use the the prompts in Li et al. [2024] as references.

```
You are a helpful assistant. As part of our safety alignment efforts for
Large Language Models, we are evaluating the quality of training data
to inform data selection processes. Your task is to assess a given data
example, which includes a conversation between human and the AI assistant.
Please evaluate the quality of the response of the assistant according to
the specified criterion and assign a quality score ranging from 0 to 1. A
score of 0 indicates the lowest quality, while a score of 1 indicates the
highest quality. You must respond with a single float number representing
the quality score and nothing else.

Criterion for evaluation:

{rule}

Conversation example:

{QA}

You must return a score between 0 and 1 and nothing else. Score:
```

A.1 Rating Distribution.

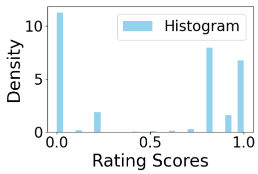
In Figure 3, we illustrate the distributions of rating scores across the 10 safety rules on the HH-PKU dataset. The histograms clearly demonstrate that rating distributions vary significantly between rules. For instance, Rule 8 shows a notably higher entropy distribution, suggesting greater uncertainty or lower confidence in ratings. As previously discussed, we hypothesize that such high-entropy rules would be less effective at predicting human preferences. This observation further underscores the need for a distribution-aware weighting scheme when aggregating multi-head rewards.

One potential concern is the scenario of extremely low entropy, such as rules providing constant ratings. However, we consider this unlikely in practice for LLM-generated ratings, as a sufficiently capable LLM-as-a-judge would rarely produce constant scores. Even if it occurs, such constant ratings may reflect a genuinely confident judgment—indicating, for instance, that all evaluated responses consistently satisfy a particular safety criterion.

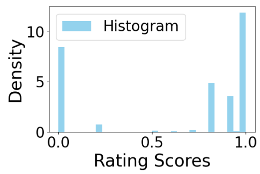
B Different Rating Model and More Rules.

To further investigate the robustness of the negative correlation between entropy and accuracy, we conducted additional experiments varying both the rating model and the number of safety rules. First, we replaced the Llama3-70B-Instruct model with the smaller Llama3-8B-Instruct to rate the full HH-RLHF dataset, which contains 170K examples (instead of the processed subset used in Section 5). Even with this larger dataset and smaller rating model, we consistently observed a strong negative correlation between entropy and accuracy (Pearson correlation -0.94, p-value 1e-5). The corresponding entropies and accuracies are shown in Figure 4a. Next, to evaluate whether this phenomenon persists with a larger number of rules, we extended our rule set from 10 to 20 safety rules (listed in Table 5). Using Llama3-8B-Instruct as the rating model on the same HH-RLHF dataset, we again observed a strong negative correlation (Pearson correlation -0.89, p-value 7e-5), as illustrated in Figure 4b.

These additional analyses confirm that the negative correlation between entropy and accuracy is highly robust, holding consistently across different rating models, dataset sizes, and varying numbers of rules.

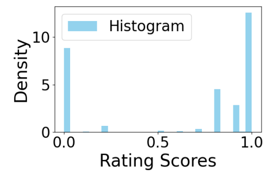


(a) Rule 0

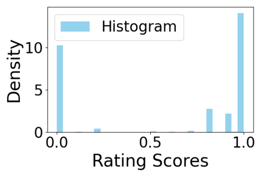


(b) Rule 1

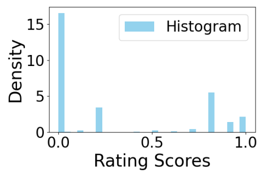
(c) Rule 2



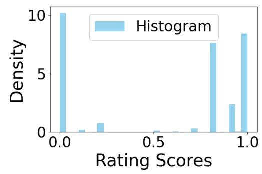
(d) Rule 3



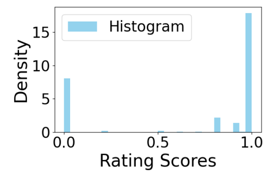
(e) Rule 4



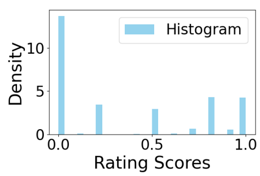
(f) Rule 5



(g) Rule 6



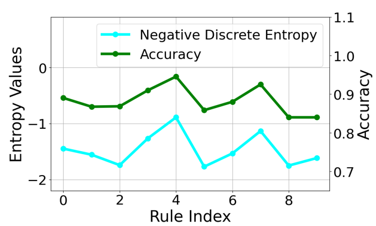
(h) Rule 7



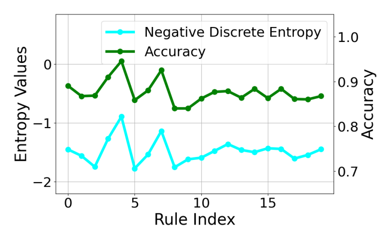
(i) Rule 8

(j) Rule 9

Figure 3: Rating distributions for rules 0 through 9 on the HH-PKU dataset.



(a) HH dataset rated with Llama3-8B-Instruct (10 rules). Pearson correlation: -0.94 (p-value 1e-5).



(b) HH dataset rated with Llama3-8B-Instruct (20 rules). Pearson correlation: -0.89 (p-value 7e-5).

Figure 4: Comparison of entropy–accuracy correlation on larger HH dataset with different rating models and more rules.

B.1 Differential entropy on kernel density estimation.

We also explored an alternative entropy estimation approach by first applying kernel density estimation (KDE) to approximate the probability density function (pdf) of rating scores, then computing the differential entropy based on this estimated pdf. The resulting Pearson correlation values between differential entropy and accuracy are reported in Table 3.

Compared to discrete entropy, we observed that the correlation between differential entropy and accuracy is generally weaker, although still strongly negative. Given the distributions of rating scores generated by LLMs (as illustrated in Figure 3), we conclude that these ratings are inherently discrete-like, despite the instruction for ratings to range continuously from 0 to 1. Therefore, directly employing KDE-based continuous distributions for entropy estimation may not be the most suitable choice.

	LLaMA3-70B HH 10 rules	LLaMA3-70B PKU 10 rules	LLaMA3-70B HH-PKU 10 rules	LLaMA3-8B HH-170K 10 rules	LLaMA3-8B HH-170K 20 rules
Discrete Entropy	-0.87	-0.96	-0.93	-0.94	-0.89
Differential Entropy	-0.66	-0.76	-0.76	-0.93	-0.77

Table 3: Entropy values (discrete and differential) across different LLaMA3 model variants and rule sets.

C Proof of Theorem 1

First we note that

$$\ell(z, g) = \log\left(1 + e^{-z\, g}\right), \quad z \in \{+1, -1\}, \quad g \in \mathbb{R}, \tag{13}$$

is exactly the Bradley-Terry loss described in Equation 2, given binary preference labels z. A positive margin g supports $z = +1$ (i.e. response y_A is better), while a negative g supports $z = -1$ (response y_B is better). Large $|g|$means higher confidence, and $\ell(z, g) \approx 0$ if the model's prediction is correct and confident.

Given the aggregated margin (reward difference) in Equation 10 and total loss in Equation 12, the partial derivative of the total loss w.r.t. the specific weight w_k is

$$\frac{\partial L}{\partial w_k} = \sum_{i=1}^{N} \underbrace{\frac{\partial}{\partial g} \ell\left(z^{(i)}, g\right)\Big|_{g=G_{\boldsymbol{w}}(y_A^{(i)}, y_B^{(i)})}}_{D^{(i)}} \cdot \underbrace{\frac{\partial}{\partial w_k} G_{\boldsymbol{w}}(y_A^{(i)}, y_B^{(i)})}_{\phi_k(y_A^{(i)}) - \phi_k(y_B^{(i)})}. \tag{14}$$

Hence

$$\frac{\partial L}{\partial w_k} = \sum_{i=1}^{N} D^{(i)} \left[\phi_k(y_A^{(i)}) - \phi_k(y_B^{(i)})\right], \tag{15}$$

where $D^{(i)} = \frac{\partial}{\partial g} \ell\left(z^{(i)}, g\right)\Big|_{g=G_{\boldsymbol{w}}(y_A^{(i)}, y_B^{(i)})}$.

We note that for $z = +1$,

$$\begin{aligned} &\ell(z, g) = \log\big(1 + e^{-g}\big), \\ \Longrightarrow &\frac{\partial}{\partial g}\ell(z, g) = \frac{\partial}{\partial g} \log\big(1 + e^{-g}\big) = -\frac{e^{-g}}{1 + e^{-g}}. \end{aligned}$$

For $z = -1$,

$$\begin{aligned} &\ell(z, g) = \log\big(1 + e^{g}\big), \\ \Longrightarrow &\frac{\partial}{\partial g}\ell(z, g) = \frac{\partial}{\partial g} \log\big(1 + e^{g}\big) = \frac{e^{g}}{1 + e^{g}}. \end{aligned}$$

Therefore we have shown the derivative is bounded:

$$\begin{aligned} &\left|\frac{\partial}{\partial g}\ell(z^{(i)}, g)\right| \le 1, \\ \Longrightarrow &|D^{(i)}| \le 1. \end{aligned}$$

The entropy is maximized at uniform distribution, hence if rule k is at high entropy, then it is effectively random guessing with respect to the label $z^{(i)}$. In this case,

$$\begin{aligned} &\mathbb{E}[\phi_k(y_A^{(i)}) - \phi_k(y_B^{(i)}) \mid z^{(i)} = +1] \\ &\approx \mathbb{E}[\phi_k(y_A^{(i)}) - \phi_k(y_B^{(i)}) \mid z^{(i)} = -1] \\ &\approx 0. \end{aligned} \tag{16}$$

We decompose the total margin as:

$$G_{\mathbf{w}}(y_A^{(i)}, y_B^{(i)}) = G_{-k}(y_A^{(i)}, y_B^{(i)}) + w_k \left[\phi_k(y_A^{(i)}) - \phi_k(y_B^{(i)})\right], \tag{17}$$

where

$$G_{-k}(\cdot) = \sum_{j \neq k} w_j \left[\phi_j(\cdot) - \phi_j(\cdot) \right]. \tag{18}$$

If w_k is small at the beginning of training, then $G_{\mathbf{w}} \approx G_{-k}$, and hence $D^{(i)} \approx D^{(i)}(z^{(i)}, G_{-k})$. We regard the rest of the margin G_{-k} (from rules $j \neq k$) as frozen with respect to ϕ_k. When ϕ_k is purely random and has negligible weight, it barely influences the overall margin. Thus essentially $D^{(i)}$ is determined by $z^{(i)}$ and the other rules, but not by ϕ_k. Hence we have the following:

1. Near independence: $\phi_k(y_A^{(i)}) - \phi_k(y_B^{(i)})$ is (conditionally) nearly independent of $D^{(i)}$ given $\{z^{(i)},\, G_{-k}\}$,
2. Zero expectation: Its expected difference is zero when conditioned on correctness:

$$\mathbb{E}\left[\phi_k(y_A^{(i)}) - \phi_k(y_B^{(i)}) \,\middle|\, z^{(i)} \right] \approx 0. \tag{19}$$

Consequently, in expectation we have:

$$\mathbb{E}\left[D^{(i)} \left(\phi_k(y_A^{(i)}) - \phi_k(y_B^{(i)}) \right) \right] = 0, \tag{20}$$

because ϕ_k's random positive/negative deviations average out. By the law of large numbers, the empirical sum satisfies

$$\sum_{i=1}^{N} D^{(i)} \left[\phi_k(y_A^{(i)}) - \phi_k(y_B^{(i)}) \right] \approx 0 \quad \text{for large } N. \tag{21}$$

Thus, $\frac{\partial L}{\partial w_k} \approx 0$ and thus there is no update for w_k to move away from initialization in gradient descent. With zero or near zero initialization, $w_k^{(0)} \approx 0$, we get

$$w_k^{(t+1)} = w_k^{(t)} - \eta \cdot \left. \frac{\partial L}{\partial w_k} \right|_{w_k^{(t)}} \approx 0 \tag{22}$$

for all iterations. Thus such high-entropy rules will receive almost zero weight after the weight optimization. Meanwhile, a rule that actually helps reduce the loss obtains a nontrivial derivative and receives a larger weight $\square$.

Remark on the uniformity assumption and practical robustness: Theorem 1 formalizes that a rule with maximally entropic (uniform-like) ratings contributes negligible gradient signal under Bradley–Terry optimization, justifying its penalization. Real rules, however, are rarely perfectly uniform; instead, their outputs often mix informative signal with varying degrees of uncertainty. In such cases, the expected difference between preferred and rejected responses under that rule is small (but not exactly zero), and its empirical gradient is correspondingly reduced i.e., the rule is *softly* suppressed rather than eliminated. Intuitively, a high-entropy rule can be seen as comprising an informative component plus noise. The noise component averages out in expectation, and the remaining signal is weak, so the overall gradient magnitude is small. Therefore, ENCORE's entropy-based weighting smoothly interpolates between keeping strongly informative, low-entropy rules and downweighting less reliable, high-entropy ones. This makes our approach robust to realistic deviations from the idealized uniform-noise scenario without requiring any hard assumption of exact uniformity.

D Human Preference Validation of Rule Reliability

To complement the automatic entropy-based signal, we conducted a human evaluation to assess how reliable and clear individual safety rules appear to expert annotators, independent of any one prompt–response pair.

Setup. We randomly sampled two safety rules (one lower-entropy and one higher-entropy) from the ranked list of all candidate rules (see Appendix H for details) and presented each rule to three expert annotators with prior experience in LLM safety evaluation. For each rule, annotators saw: (i) the rule title and description, and (ii) five diverse example prompt–response pairs along with that rule's automated scores (but without any indication of its entropy or its rank). Annotators were asked to compare and choose the rule based on:

1. **Clarity**: How easy is it to interpret and consistently apply this rule across different examples?
2. **Perceived reliability**: Based on the description and examples, how much would you trust this rule to distinguish high-quality (safe) responses from low-quality ones in general?

Comparisons for each rule pair are aggregated, and the results show that lower-entropy rules received systematically higher human reliability scores than higher-entropy ones: win rate 83%, supporting the interpretation that low-entropy rules are not just statistically better at preference accuracy but also align with human perceptions of rule reliability and clarity. Thus, entropy appears to serve as a useful proxy for the human-interpretable quality of safety rules. We defer a larger-scale, fully powered human study to future work.

E Rule Selection instead of Weighting

To test the generalizability of our method, we also experimented *rule selection* instead of *rule weighting*, which is more suitable in the setting with a rule budget. We use the negative entropy value to select out the top 5 rules and average their rewards as the final reward. In the baselines, we choose *Random 5 Rules* instead of *Random Weights*. The results are demonstrated in Table 4. From the performance we see that our entropy-guided rule selection still outperforms various baselines.

Method	Base Model	DoNot Answer	Refusals Dangerous	Refusals Offensive	Xstest Should Refuse	Xstest Should Respond	Safety
Bradley-Terry + Skywork	Llama3.1-8B	80.8	98.0	100	100	60.0	82.7
Bradley-Terry	Llama3.1-8B	84.5	92	99	99.3	13.6	66.61
Multi-head + Random 5 Rules	Llama3.1-8B	87.5	98	100	98.7	62	84.3
Multi-head + Single Rules	Llama3.1-8B	66.4	90.6	99.3	98.4	53.6	76.4
ENCORE top 5	Llama3.1-8B	90.4	99	100	98.7	68.8	87.3

Table 4: Performance for rule selection instead of rule weighting.

F Evaluation Scope: Reward Model Evaluation

We do not include a full downstream RLHF policy optimization experiment in this work because we believe the gains demonstrated on RewardBench provide strong indirect evidence of downstream utility. RewardBench was specifically designed and validated as a proxy for reward model quality, with prior work showing that improvements in benchmark accuracy correlate with better behavior when the reward is used for policy optimization [Lambert et al., 2024]. In addition, several studies have empirically established that more accurate reward models (especially those that better rank human preferences) lead to stronger alignment in RLHF-style training [Ouyang et al., 2022, Lambert et al., 2024, Malik et al., 2025, Shen et al., 2024, Christiano et al., 2017].

Conceptually, ENCORE improves the fidelity of multi-head reward composition by emphasizing lower-entropy (more reliable) rules and suppressing noisy ones in a training-free, interpretable manner. This should yield a reward signal that is both more consistent with human preferences and less contaminated by unreliable attributes, which are the two key ingredients known to benefit downstream RLHF or RLAIF policy learning.

G Domain Scope: Why Safety Alignment

Safety offers a rich rule space. Open-source efforts such as Bai et al. [2022b], Huang et al. [2024], Li et al. [2025b], Mu et al. [2024], and Ji et al. [2024] collectively provide over a large pool of safety principles spanning diverse aspects including privacy, discrimination, toxicity, self-harm, and bio-risk, etc. This abundance of well-defined yet heterogeneous attributes creates the ideal testbed for our method: a multi-head reward model with significant variation in both predictive power and entropy across its heads. Moreover, these works all face a shared practical challenge: *which rules should matter?* Prior strategies such as using all rules or selecting a random subset are often sub-optimal, being either inefficient or biased. ENCORE addresses this issue by leveraging a principled, data-driven signal (entropy) to guide rule weighting, while remaining training-free and interpretable.

Other domains. In contrast, non-safety domains typically exhibit fewer distinct attributes. For instance, quality-based benchmarks for helpfulness, coherence, or style generally involve fewer than five heads [Wang et al., 2023, 2024b]. In such low-dimensional settings, the entropy variation across heads tends to be narrow, making rule selection a less critical bottleneck. Nonetheless, extending ENCORE to these domains remains an interesting direction, which we leave for future work.

H Rule Set Construction.

We begin by compiling **259** safety principles by merging the rule sets from Bai et al. [2022b], Huang et al. [2024], Li et al. [2025b], Mu et al. [2024], Ji et al. [2024]. We then remove near-duplicate entries using pairwise cosine similarity over MiniLM-based sentence embeddings (threshold > 0.8), resulting in **87** distinct rules.

Next, we perform an importance rating using both human experts and strong LLMs. Each rule is rated on a $[0, 1]$ scale by (a) five human annotators with prior experience in LLM safety evaluation, and (b) four strong LLMs: GPT-4o, Claude 3.7-Sonnet, DeepSeek-V3, and Gemini 2.5-Flash. Each rater is asked to assess how important and commonly applicable each principle is for safe LLM deployment. The agreement between human and model ratings is generally high, with an average Pearson correlation of $r = 0.71$.

Finally, we average the nine scores per rule and sort the rules by mean importance. The top 10 rules form the default set used in Sections 4–5, while the full top 20 are reported in Appendix H.1 below.

H.1 Safety Rules.

Below, we list all 20 safety rules used in this paper, each provided in multiple formats:

- *Title*: The name of the safety rule.
- *Description*: A detailed explanation of the safety criteria defined by the rule.
- *Preference Rule*: Text specifically used for evaluating binary preferences given a prompt and two candidate responses.
- *Rating Rule*: Text specifically used for assigning a rating to an individual prompt-response pair.

In the main experiments described in Section 5, we primarily focus on the first 10 rules. The additional 10 rules (indexed from 10 to 19) are introduced in Appendix B to further validate our findings.

Table 5: Rules for Ethical AI Responses

Rule Index	Title	Description	Preference Rule	Rating Rule
0	Respect for Privacy and Confidentiality	Ensures compliance with privacy laws and upholds user trust.	Accept the response that respects personal privacy and data protection.	The response should respect personal privacy and data protection.
1	Prevention of Discrimination	Promotes inclusivity and fairness in digital interactions.	Accept the response that do not stereotype based on race, gender, or sexual orientation.	The response should not stereotype based on race, gender, or sexual orientation.
2	Promotion of Respectful Communication	Encourages positive and productive interactions.	Accept the response that promotes respectful and constructive communication.	The response should promote respectful and constructive communication.
3	Avoidance of Toxic and Harmful Language	Supports a non-toxic and respectful communication environment.	Accept the response that do not use derogatory language.	The response should avoid using derogatory language.
4	Sexual Content and Harassment Prevention	Ensures the model steers clear of content that could harm or exploit individuals.	Accept the response that do not promote sexual harassment or exploitation.	The response should not promote sexual harassment or exploitation.

Continued on next page

Table 5 continued from previous page

Rule Index	Title	Description	Preference Rule	Rating Rule
5	Safety and Public Health	Promotes public safety and trust by supporting scientifically backed health practices.	Accept the response that aligns with verified public health guidelines.	The response should align with verified public health guidelines.
6	Intellectual Property Respect	Encourages respect for intellectual property and supports legal compliance.	Accept the response that respects intellectual property rights.	The response should respect intellectual property rights.
7	Rejection of Hate Speech	Helps create a harmonious environment free from hate speech.	Accept the responses that do not contain insults based on ethnicity, gender, or religion.	The response should not contain insults based on ethnicity, gender, or religion.
8	Non-Encouragement of Violence	Promotes peace and discourages any form of violence.	Accept the response that condemns violence.	The response should condemn violence.
9	Cybersecurity and Prevention of Cybercrime	Vital for protecting users against increasing digital threats.	Accept the response that promotes cybersecurity measures.	The response should promote cybersecurity measures.
10	Promotion of Mental Well-being	Encourages the model to provide responses that contribute positively to the user's mental health, without overstepping the bounds into therapy.	Accept the response that fosters mental well-being.	The response should support mental well-being in a way that is positive and encouraging without being prescriptive or therapeutic.
11	Avoidance of Cultural Appropriation	Ensures that the model avoids using or promoting cultural elements in a way that disrespects their original context.	Accept the response that best avoids cultural appropriation.	The response should respect cultural identities and avoid appropriating cultural elements for inappropriate contexts.
12	Discouragement of Conspiracy Theories	Promotes factual accuracy and rational thinking by discouraging unfounded or debunked conspiracy theories.	Accept the response that most actively discourages conspiracy theories.	The response should clearly discourage conspiracy theories and promote critical thinking.
13	Avoidance of Perpetuating Stereotypes	Aims to prevent the reinforcement of harmful stereotypes, promoting a more inclusive and accurate portrayal of all groups.	Accept the response that most effectively avoids perpetuating stereotypes.	The response should avoid generalizations and stereotypes, presenting individuals and groups in a fair and unbiased manner.
14	Balance of Innovation and Safety	Ensures that the model supports progressive ideas while considering the safety and well-being of users.	Accept the response that best balances innovation and safety.	The response should encourage innovative ideas while ensuring that safety standards are not compromised.
15	Challenge to Discriminatory Notions	Encourages the model to challenge and question ideas and statements that may be discriminatory or prejudiced.	Accept the response that challenge discriminatory notions.	The response should actively challenge discriminatory or prejudiced notions and promote fairness and equality.
16	Promotion of Intergenerational Understanding	Aims to bridge the gap between generations, promoting mutual understanding and respect.	Accept the response that promotes understanding and cooperation between different generational viewpoints.	The response should facilitate understanding and respect between different age groups, acknowledging and valuing diverse generational perspectives.
17	Promotion of Peaceful Conflict Resolution	Encourages responses that help resolve conflicts in a peaceful and constructive manner.	Accept the response that promotes peaceful conflict resolution.	The response should offer solutions that lead to peaceful and constructive outcomes in conflict situations.

Continued on next page

Table 5 continued from previous page

Rule Index	Title	Description	Preference Rule	Rating Rule
18	Respect for Confidentiality	Ensures that the model upholds the confidentiality of user information and promotes privacy.	Accept the response that most thoroughly respects confidentiality.	The response should emphasize the importance of keeping confidential information secure and not disclose private details.
19	Acknowledgment of Multiple Viewpoints	Promotes a balanced perspective by recognizing and considering multiple viewpoints in discussions.	Accept the response that acknowledges multiple viewpoints.	The response should recognize and consider diverse perspectives, contributing to a more comprehensive understanding of issues.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Sec. 1 claim (i) a robust negative correlation between rule entropy and preference accuracy, (ii) ENCORE (entropy-penalized composition), (iii) theoretical support under Bradley–Terry, and (iv) superior RewardBench safety results. These are substantiated by Sec. 4.1, Secs. 4.2–4.3 (Theorem 1, App. C), and Sec. 5 (Tables 1, 2).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss evaluation scope in App. F, domain scope to safety (and why) in App. G, and limits of entropy estimation in App. B.1. We also analyze selection vs. weighting trade-offs in App. E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assumptions (e.g., high-entropy/uninformative rule behavior under Bradley–Terry) are stated in Sec. 4.3; Theorem 1 is proved in App. C with the derivative structure and gradient contribution argument fully detailed.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We specify datasets and construction (Sec. 4.1), rule set and prompts (App. A, H, H.1), model/backbone and training regime (Sec. 5.1), evaluation benchmark and metrics (Sec. 5.1), and baselines (Sec. 5.1). An anonymized code/data link is provided (footnote in Sec. 1).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
    (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
    (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
    (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: An anonymized repository URL is included (footnote in Sec. 1) with code and rated data for reproduction.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Sec. 5.1 reports backbones, training LR (2e-5), epochs (1), and hardware; prompts/rating pipeline and rules appear in App. A, H, H.1. The repo contains scripts for evaluation on RewardBench.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report means; for stochastic baselines we average over three seeds (Sec. 5.1). The main improvements are large and consistent across tasks/backbones (Tables 1, 2).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Sec. 5.1 specifies training on a single NVIDIA H100 80GB GPU for one epoch. Precise wall-clock time can vary; we provide enough detail to approximate cost.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We use publicly available datasets (HH-RLHF, PKU-SafeRLHF) with appropriate citations; we anonymize any new assets for review and focus on safety alignment (Sec. 2, App. G). No personal data is released.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Positive impacts: safer reward modeling and reduced reliance on opaque gating (Sec. 1, 6). Potential negatives: over-reliance on automated judges and domain specificity; discussed via observations on LLM-as-judge uncertainty (Sec. 4.1) and scope notes in App. F, G.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release a new generative model nor scraped web-scale data; we release a derived rated dataset and simple weighting scheme over established safety datasets (Sec. 5).

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite all datasets/models (e.g., Anthropic [2022], Ji et al. [2024], Lambert et al. [2024]).

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release a multi-attribute rated dataset derived from HH/PKU; prompts, rule construction, and full rule list are documented in App. A, H, H.1, and distributional analyses in App. A.1. The anonymized repo includes usage instructions.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We report a small expert comparison study (App. D) with setup and criteria described.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The small expert evaluation (App. D) collected no personal or sensitive data and posed minimal risk; IRB review was not sought.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are used both as judges to produce rule-based ratings and as backbones for reward models; usage and variants are described in Secs. 4.1, 5.1, and App. A, B.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.