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

Xiaomin Li[1*], **Xupeng Chen**[2], **Jingxuan Fan**[1], **Eric Hanchen Jiang**[3], **Mingye Gao**[4]
[1]Harvard University
[2]New York University
[3]University of California, Los Angeles
[4]Massachusetts Institute of Technology

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.

Code & Data — https://github.com/XiaominLi1998/Submission-ENCORE

Extended version — https://arxiv.org/abs/2503.20995

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 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

*Correspondence to: Xiaomin Li (xiaominli@g.harvard.edu).

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.

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 a prompt and two candidate responses,

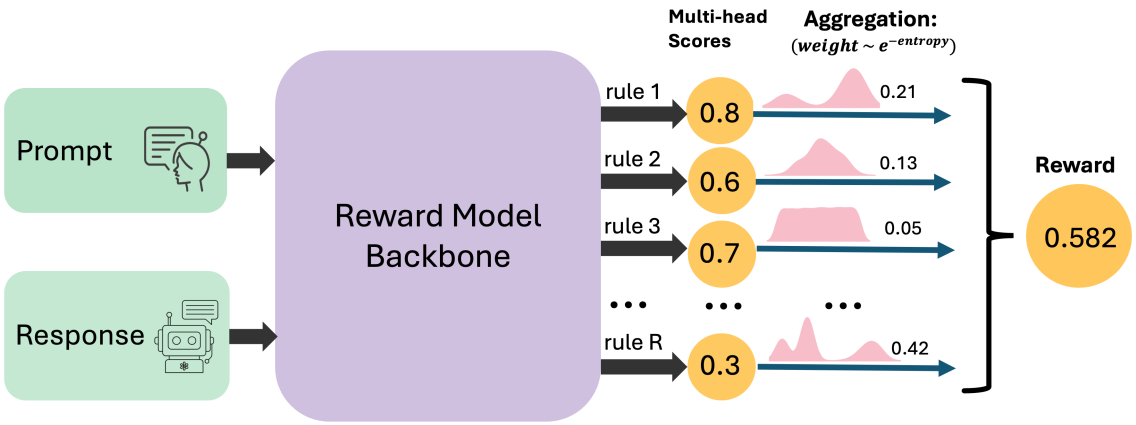


Figure 1: Pipeline of our ENCORE framework. Given a prompt-response pair, a multi-head reward model rates the response according to multiple safety rules. Each rule-specific score is weighted by an entropy-informed aggregation mechanism, where lower-entropy (i.e., more reliable) rules are assigned higher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

Bradley-Terry models the probability that response y_A is preferred over y_B as

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

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 $\hat{\mathcal{H}}(\psi_k)$.[1]

[1]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

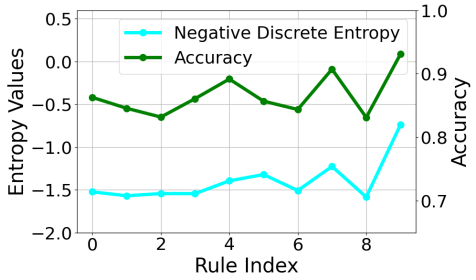
We begin with preliminary experiments to empirically validate the relationship between the entropy of rule-specific ratings and their accuracy in distinguishing human-preferred responses.

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:

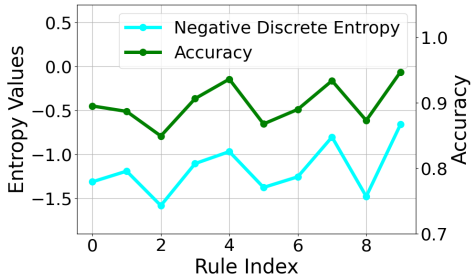
> **Title:** Prevention of Discrimination
> **Description:** Promotes inclusivity and fairness in digital interactions.
> **Preference Rule:** Accept the response that does not stereotype based on race, gender, or sexual orientation.
> **Rating Rule:** The response should not stereotype based on race,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

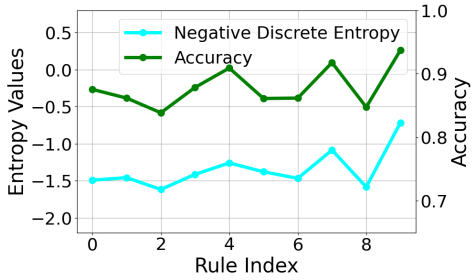
Correlation between Entropy and Accuracy. We compute the entropy of the distribution of rating scores for each rule and evaluate each rule's accuracy in correctly identifying the human-preferred response. Figure 2 illustrates the clear, consistent negative correlation between (negative) entropy and accuracy 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 C). 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.



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

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

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

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 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 E.

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

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

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.

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.

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

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

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

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. Addition-

ally, 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 G, this simpler approach still outperforms random selection 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. 2024. Reward-Bench: A Comprehensive Benchmark for Reward Models. https://huggingface.co/spaces/allenai/reward-bench.

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

Anthropic. 2024. Introducing Claude 3.5 Sonnet. Introduces Claude 3.5 Sonnet with improved performance in intelligence, vision capabilities, and new Artifacts feature.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bai, Y.; Kadavath, S.; Kundu, S.; Askell, A.; Kernion, J.; Jones, A.; Chen, A.; Goldie, A.; Mirhoseini, A.; McKinnon, C.; et al. 2022b. Constitutional AI: harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Dorka, N. 2024. Quantile Regression for Distributional Reward Models in RLHF. *arXiv preprint arXiv:2409.10164*.

Du, N.; Huang, Y.; Dai, A. M.; Tong, S.; Lepikhin, D.; Xu, Y.; Krikun, M.; Zhou, Y.; Yu, A. W.; Firat, O.; et al. 2022. GLaM: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, 5547–5569. PMLR.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ganguli, D.; Askell, A.; Schiefer, N.; Liao, T. I.; Lukošiūtė, K.; Chen, A.; Goldie, A.; Mirhoseini, A.; Olsson, C.; Hernandez, D.; et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Glaese, A.; McAleese, N.; Trebacz, M.; Aslanides, J.; Firoiu, V.; Ewalds, T.; Rauh, M.; Weidinger, L.; Chadwick, M.; Thacker, P.; et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Huang, S.; Siddarth, D.; Lovitt, L.; Liao, T. I.; Durmus, E.; Tamkin, A.; and Ganguli, D. 2024. Collective Constitutional AI: Aligning a Language Model with Public Input. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1395–1417.

Ji, J.; Hong, D.; Zhang, B.; Chen, B.; Dai, J.; Zheng, B.; Qiu, T.; Li, B.; and Yang, Y. 2024. PKU-SafeRLHF: Towards Multi-Level Safety Alignment for LLMs with Human Preference. *arXiv preprint arXiv:2406.15513*.

Kundu, S.; Bai, Y.; Kadavath, S.; Askell, A.; Callahan, A.; Chen, A.; Goldie, A.; Balwit, A.; Mirhoseini, A.; McLean, B.; et al. 2023. Specific versus general principles for Constitutional AI. *arXiv preprint arXiv:2310.13798*.

Lambert, N.; Pyatkin, V.; Morrison, J.; Miranda, L.; Lin, B. Y.; Chandu, K.; Dziri, N.; Kumar, S.; Zick, T.; Choi, Y.; et al. 2024. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.

Lee, H.; Phatale, S.; Mansoor, H.; Mesnard, T.; Ferret, J.; Lu, K.; Bishop, C.; Hall, E.; Carbune, V.; Rastogi, A.; and Prakash, S. 2025. RLAIF vs. RLHF: scaling reinforcement learning from human feedback with AI feedback. In *International Conference on Machine Learning*. PMLR.

Li, X.; Gao, M.; Zhang, Z.; Fan, J.; and Li, W. 2025a. Data-adaptive Safety Rules for Training Reward Models. *arXiv preprint arXiv:2501.15453*.

Li, X.; Gao, M.; Zhang, Z.; Fan, J.; and Li, W. 2025b. RuleAdapter: Dynamic Rules for training Safety Reward Models in RLHF. In *Forty-second International Conference on Machine Learning*.

Li, X.; Gao, M.; Zhang, Z.; Yue, C.; and Hu, H. 2024. Rule-based data selection for large language models. *arXiv preprint arXiv:2410.04715*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, C. Y.; Zeng, L.; Liu, J.; Yan, R.; He, J.; Wang, C.; Yan, S.; Liu, Y.; and Zhou, Y. 2024b. Skywork-reward: Bag of tricks for reward modeling in LLMs. *arXiv preprint arXiv:2410.18451*.

Malik, S.; Pyatkin, V.; Land, S.; Morrison, J.; Smith, N. A.; Hajishirzi, H.; and Lambert, N. 2025. RewardBench 2: Advancing Reward Model Evaluation. *arXiv preprint arXiv:2506.01937*.

Mu, T.; Helyar, A.; Heidecke, J.; Achiam, J.; Vallone, A.; Kivlichan, I. D.; Lin, M.; Beutel, A.; Schulman, J.; and Weng, L. 2024. Rule Based Rewards for Language Model Safety. *Advances in Neural Information Processing Systems*, 37.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36: 53728–53741.

Ramamurthy, R.; Ammanabrolu, P.; Brantley, K.; Hessel, J.; Sifa, R.; Bauckhage, C.; Hajishirzi, H.; and Choi, Y. 2022. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shen, W.; Zhang, X.; Yao, Y.; Zheng, R.; Guo, H.; and Liu, Y. 2024. Improving reinforcement learning from human feedback using contrastive rewards. *arXiv preprint arXiv:2403.07708*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wang, H.; Xiong, W.; Xie, T.; Zhao, H.; and Zhang, T. 2024a. Interpretable Preferences via Multi-Objective Reward Modeling and Mixture-of-Experts. *arXiv preprint arXiv:2406.12845*.

Wang, Z.; Dong, Y.; Delalleau, O.; Zeng, J.; Shen, G.; Egert, D.; Zhang, J. J.; Sreedhar, M. N.; and Kuchaiev, O. 2024b. HelpSteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*.

Wang, Z.; Dong, Y.; Zeng, J.; Adams, V.; Sreedhar, M. N.; Egert, D.; Delalleau, O.; Scowcroft, J. P.; Kant, N.; Swope, A.; et al. 2023. HelpSteer: Multi-attribute helpfulness dataset for SteerLM. *arXiv preprint arXiv:2311.09528*.

Wu, Z.; Hu, Y.; Shi, W.; Dziri, N.; Suhr, A.; Ammanabrolu, P.; Smith, N. A.; Ostendorf, M.; and Hajishirzi, H. 2023. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.