Enhancing Reinforcement Learning with Label-Sensitive Reward for Natural Language Understanding

Anonymous ACL submission

Abstract

Recent strides in large language models (LLMs) have yielded remarkable performance, leveraging reinforcement learning from human feedback (RLHF) to significantly enhance generation and alignment capabilities. However, RLHF encounters numerous challenges, including the *objective mismatch* issue, leading to suboptimal performance in Natural Language Understanding (NLU) tasks. To address this limitation, we propose a novel Reinforcement Learning framework enhanced with Label-sensitive Reward (RLLR) to amplify the performance of LLMs in NLU tasks. By incorporating label-sensitive pairs into reinforcement learning, our method aims to adeptly capture nuanced label-sensitive semantic features during RL, thereby enhancing natural language understanding. Experiments conducted on five diverse foundation models across eight tasks showcase promising results. In comparison to Supervised Fine-tuning models (SFT), RLLR demonstrates an average performance improvement of 1.54%. Compared with RLHF models, the improvement averages at 0.69%. These results reveal the effectiveness of our method for LLMs in NLU tasks.

1 Introduction

Large language models (LLMs) (Achiam et al., 2023; Chowdhery et al., 2023; Touvron et al., 2023a) have undergone impressive advancements which transform NLP tasks into a unified text-to-text paradigm, achieving robust alignment and generation capabilities through reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022; Bai et al., 2022a). Particularly, models are required to predict the correct labels in natural language understanding (NLU) tasks, distinct from natural language generation (NLG) tasks. Numerous studies have employed rationales to assist LLMs with Chain-of-Thought (CoT) prompting during supervised fine-tuning (SFT) stage (Kim et al., 2023; Hsieh et al., 2023).

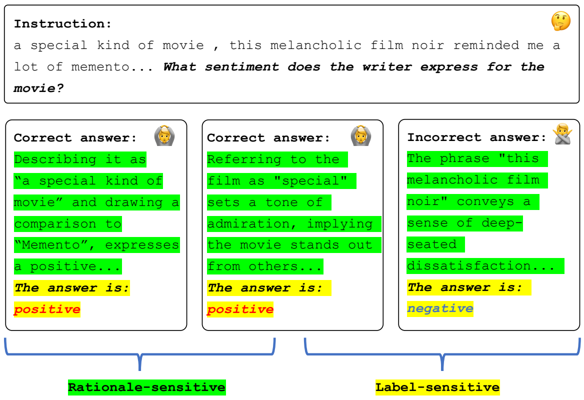

Figure 1: The example of rationale-sensitive and label-sensitive pairs from sentiment classification. Highlight rationales in green and labels in yellow.

However, Lambert and Calandra (2023) detail a fundamental challenge in RLHF learning schemes: the *objective mismatch* issue. This arises when the reward model is influenced by human preference data, introducing biases that conflict with downstream evaluation metrics, especially when applied to NLU tasks. In RLHF, comparison data is initially sampled from the SFT model and ranked by a labeler. Then the policy model is optimized against the reward model that is trained with these pairs to align with human preference. For NLU tasks, the pairs can be categorized into rationale-sensitive and label-sensitive. Figure 1 illustrates three answers sampled from the SFT model for one instruction. If two answers have the same label and different rationales, they form a rationale-sensitive pair, with the more reasonable rationale considered superior. In contrast, if two answers have different labels, they form a label-sensitive pair, with the correct label deemed superior. However, we observed that the pairs sampled from the SFT model mainly fall into the category of rationale-sensitive. Figure 2 shows the specific distribution ratios of pairs across several NLU tasks. The percentage of rationale-

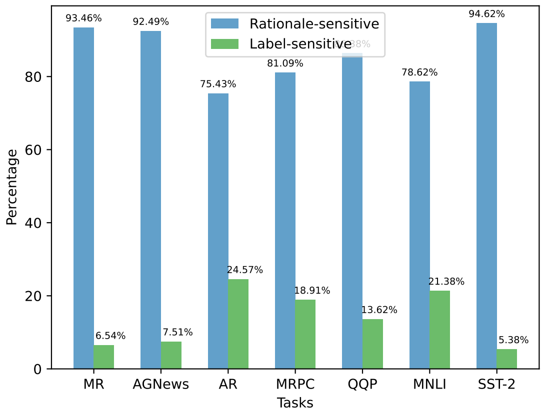

Figure 2: The distribution of rationale-sensitive and label-sensitive pairs sampled from SFT model across a range of tasks.

sensitive pairs exceeds 75%, and in datasets like SST-2, MR, and AGNews, surpasses 90%. The severe imbalance in the distribution of pairs leads the model to prioritize the quality of rationales over the correctness of labels during RLHF training, which conflicts with the evaluation metric (mostly *label accuracy*) of NLU tasks. A detailed analysis is presented in Section 4.2.

To address this challenge, our paper proposes a Reinforcement Learning framework enhanced with Label-sensitive Reward (RLLR) for NLU tasks. Firstly, we generate rationales corresponding to the gold labels of the training data. The SFT model is trained with rationales, incorporating CoT prompting to enhance comprehension abilities. Secondly, we generate rationales for the incorrect labels (relative to the gold labels). Unlike RLHF, which uses human intervention to rank sentences, RLLR automatically constructs label-sensitive pairs for training the reward model based on the correctness of the label. Finally, we train the policy model against the label-sensitive reward model with Proximal Policy Optimization (PPO) to prioritize the correctness of labels. Furthermore, optimizing with mixed rewards from the label-sensitive and rationale-sensitive reward models, $\text{RLLR}_{\text{MIXED}}$ ensures both the accuracy of labels and the quality of rationales. Extensive experiments on eight NLU tasks demonstrate that our method consistently outperforms the SFT baseline by an average of 1.54% and the RLHF baseline by an average of 0.69%, while also exhibiting higher quality in rationales generation.

Our contributions are summarized as:

(1) We propose a Reinforcement Learning framework enhanced with Label-sensitive Reward (RLLR) for NLU tasks to tackle the *objective mismatch* issue.

(2) Optimizing with mixed rewards, $\text{RLLR}_{\text{MIXED}}$ can achieve promising performance on both the accuracy of labels and the quality of rationales.

(3) Through empirical experiments, we demonstrate the effectiveness of our method. We also conduct in-depth analyses of the role of rationales, the performance of reward models, and the quality of generated rationales.

2 Related Work

Reinforcement Learning from Human Feedback. LLMs have demonstrated commendable performance, leveraging RLHF to achieve notable alignment and generation capabilities (Ouyang et al., 2022; Achiam et al., 2023; Bai et al., 2022a; Ziegler et al., 2019). RLHF aims to optimize the policy language model to generate content that is desired by humans. Recently, some research endeavors have uncovered inherent challenges in RLHF (Casper et al., 2023; Lambert et al., 2023), including feedback type limitations, evaluation difficulties, oversight challenges, etc. Several methods have been proposed to mitigate these challenges. Bai et al. (2022b) introduce RL from AI Feedback (RLAIF), training an AI assistant through self-improvement while adhering to constitutional principles that constrain model-generated content. Wu et al. (2023) introduce a fine-grained RLHF framework that uses fine-grained human feedback, such as identifying false sentences or irrelevant sub-sentences, as an explicit training signal. However, these approaches encounter a fundamental challenge in RLHF learning schemes: the *objective mismatch* issue (Lambert and Calandra, 2023). In this paper, we tackle this problem by training the reward model with the label-sensitive pairs.

Chain-of-Thought. CoT can significantly improve the complex reasoning ability of LLMs by generating natural language rationales that lead to the final answer (Wei et al., 2022; Kim et al., 2023). Hsieh et al. (2023) introduce a distilling mechanism step-by-step, extracting LLM rationales as additional supervision for training small models within a multi-task framework. Fu et al. (2023) propose a method to specialize the model's ability (smaller than 10B) towards a target task with CoT prompting. In this paper, we enhance the performance of LLMs on NLU tasks with CoT prompting utilizing rationales generated for the labels.

3 Proposed Method

In this section, we introduce the training pipeline of our method as illustrated in Figure 3, including supervised fine-tuning, reward model training, and reinforcement learning enhanced with mixed rewards.

3.1 Supervised Fine-Tuning

In NLU tasks, the supervised dataset is denoted as $\mathcal{S} = \{x, y\}$, where x denotes the sentence and y denotes the class label. The unsupervised dataset is denoted as $\mathcal{U}$, and the foundation model is denoted as π. According to Wei et al. (2022), generating rationales that lead to the final answer can significantly improve the reasoning ability of LLMs through CoT. Therefore, we first generate a rationale from the sentence x and the label y with a specific prompt template using either human annotators or LLMs such as GPT-4. Then we reform the original dataset $\mathcal{S}$ to the training dataset $\mathcal{T} = \{q, a\}$. The question q is constructed by x with a template and the answer a with t tokens is obtained by combining rationale and label, denoted as $a = a_{1,\cdots,t}$. The details of prompts can be found in Appendix B and C. The foundation model π is then trained on $\mathcal{T}$ to obtain the model π_{SFT}. Formally, the loss for supervised fine-tuning is defined as:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(q,a)\sim\mathcal{T}}\left[\log P_\pi\left(a_t \mid q, a_{1,\cdots,t-1}\right)\right]. \tag{1}$$

3.2 Reward Model Training

In the second phase, comparison data are sampled from the answers generated by the SFT model π_{SFT} given a question. As illustrated in Figure 2, more than 75% of the pairs generated by the SFT model are rationale-sensitive pairs (i.e., both answers have the same label). The sentences in the rationale-sensitive pair are then labeled with a preference order. In RL, the reward model denoted as r_ϕ assigns higher scores to preferable answers compared with unfavorable ones, employing the Bradley-Terry paired comparison (Bradley and Terry, 1952). In this case, the model pays more attention to the quality of the generated rationales rather than the correctness of the labels. This leads to suboptimal performance due to the *objective mismatch* issue mentioned earlier.

To address this issue, we generate rationales based on the incorrect label for an input sentence and combine them to form a new answer. we generate rationales for incorrect labels $\hat{y}$ to create a new answer $\hat{a}$, which is a rationale-augmented incorrect answer. Along with the correct answer a, we can obtain the preferences $a \succ \hat{a} \mid q$ for label-sensitive pairs without extra annotation. The details of label-sensitive pair construction can be found in Appendix E. The label-sensitive and rationale-sensitive pairs are used to train two reward models, respectively. Specifically, we have $a^1 \succ a^2 \mid q$ to represent the preference in the pair. To predict these preferences, we employ the Bradley-Terry (BT) model, which defines the preference probability as follows:

$$P_{BT} = \frac{\exp\left(r_\phi\left(q, a^1\right)\right)}{\exp\left(r_\phi\left(q, a^1\right)\right) + \exp\left(r_\phi\left(q, a^2\right)\right)}. \tag{2}$$

This objective is framed as a binary classification problem to train the reward model $r_\phi(q, a)$ with the loss defined as:

$$\mathcal{L}_R = -\mathbb{E}_{(q,a^1,a^2)\sim\mathcal{C}}\left[\log\sigma\left(r_\phi\left(q, a^1\right) - r_\phi\left(q, a^2\right)\right)\right], \tag{3}$$

where σ is the logistic function and $\mathcal{C}$ is the dataset of comparisons. In this way, we can obtain two separate reward models $r_{\phi1}$ and $r_{\phi2}$ with the label-sensitive and rationale-sensitive pairs, respectively. The reward model $r_\phi(q, a)$ is often initialized from the SFT model $\pi_{\text{SFT}}(a|q)$ with the addition of a linear layer on top of the final transformer layer that produces a single scalar prediction for the reward value.

3.3 Reinforcement Learning

During the RL phase, we use the reward model to train the SFT model π_{SFT} using Proximal Policy Optimization (PPO) on the unsupervised dataset $\mathcal{U}$. Given a question constructed by the sentence from $\mathcal{U}$, the mixed reward function from $r_{\phi1}(q, a)$ and $r_{\phi2}(q, a)$ is calculated as :

$$r_{\text{M}}(q, a) = \begin{cases} r_{\phi1}(q, a) + r_{\phi2}(q, a), & \text{if } r_{\phi1}(q, a) < \lambda \\ \lambda + r_{\phi2}(q, a), & \text{if } r_{\phi1}(q, a) \geq \lambda \end{cases} \tag{4}$$

where λ is a hyper-parameter as the threshold for $r_{\phi1}$ (the label-sensitive reward model). According to experimental observations, the reward score of $r_{\phi1}$ converges to around 5.0, while the score of $r_{\phi2}$ is within 1.0, resulting in an imbalance between the two. To prevent reinforcement learning from being

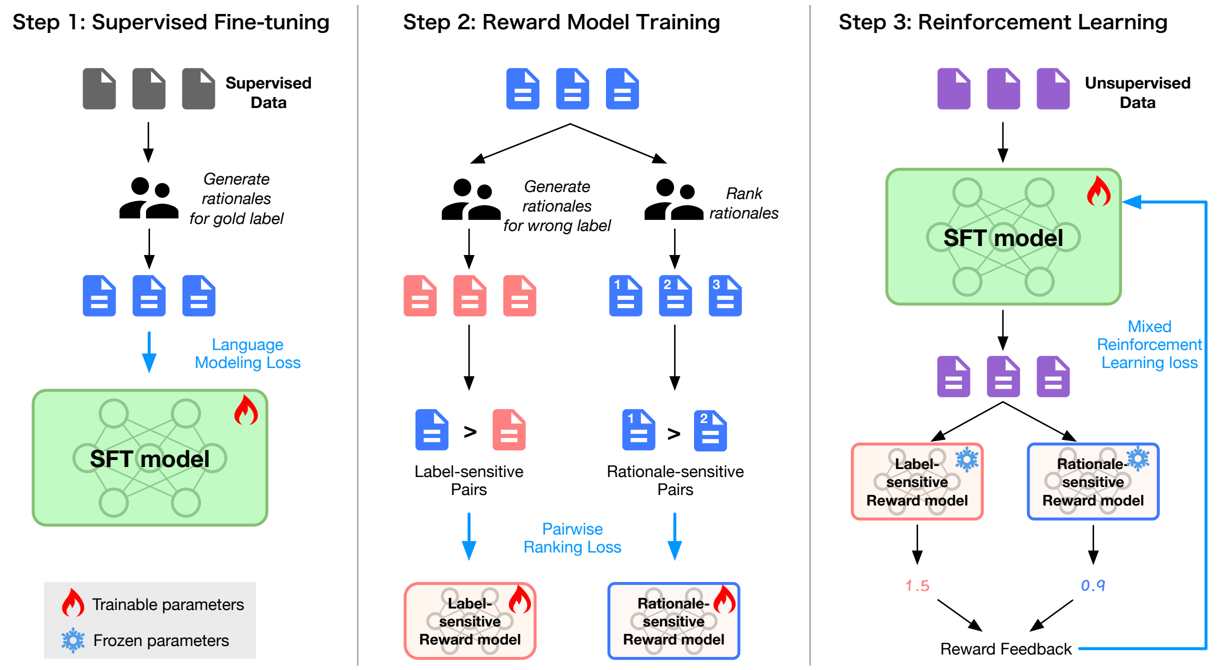

Figure 3: The training pipeline of RLLR with supervised fine-tuning, reward model training, and mixed reinforcement learning. Blue arrows indicate data used for model training.

completely dominated by $r_{\phi 1}$, we set a threshold value λ. When the score of $r_{\phi 1}$ is less than λ, the combined reward score is the sum of $r_{\phi 1}$ and $r_{\phi 2}$; when the score of $r_{\phi 1}$ is greater than or equal to λ, the combined reward score is equal to λ plus $r_{\phi 2}$. We first optimize the policy based on $r_{\phi 1}$, focusing on the correctness of the labels. As the RL training progresses, the score of $r_{\phi 1}$ gradually exceeds λ. Once the score of $r_{\phi 1}$ surpasses λ, we truncate it. At this point, the model will pay more attention to $r_{\phi 2}$, which is the quality of the rationale. In this way, both $r_{\phi 1}$ and $r_{\phi 2}$ can play a role in reinforcement learning, allowing the final policy model to predict the correct labels while generating high-quality rationales.

To guide the RL training, the loss function is constructed by combining the rewards generated by the reward model with a KL divergence constraint, which ensures that the policy does not deviate significantly from its initial behavior, defined as:

$$\max_{\pi_{\mathrm{RL}}} \mathbb{E}_{(q,a)\sim D_{\pi_{\mathrm{RL}}}} \left[r_{\mathrm{M}}(q,a) - \beta \log\left(\frac{\pi_{\mathrm{RL}}(a|q)}{\pi_{\mathrm{SFT}}(a|q)}\right)\right], \tag{5}$$

where π_{RL} is the learned RL policy, π_{SFT} is the SFT model, and β is the KL reward efficient controlling the strength of the KL penalty. RLLR$_{\mathrm{MIXED}}$ is obtained from this objective with two reward models while RLLR is trained only with the label-sensitive rationale reward model.

4 Experiments

4.1 Experiment Setup

Datasets. We evaluate the performance of our proposed method across eight NLU tasks, encompassing five from the GLUE benchmark (Wang et al., 2018). The tasks include Movie Reviews (MR) (Pang and Lee, 2005), AppReviews (AR) (Grano et al., 2017) and SST-2 for sentiment classification, AGNews (Zhang et al., 2015) for topic classification, MRPC and QQP for paraphrase detection, MNLI for textual entailment, and STS-B for semantic similarity. We employ the Pearson correlation coefficient as our evaluation metric for STS-B, and accuracy for others. To ensure a fair comparison with baseline methods, we convert all tasks into a text-to-text format following (Sanh et al., 2021). For methods that require rationales, we utilize GPT-4 to generate rationales conditioned on given labels. The details regarding the prompt templates and rationale annotation process are provided in Appendix B and C.

Baselines. We conduct our experiments using several state-of-the-art foundation models, including LLaMA2 (Touvron et al., 2023b), Baichuan2 (Yang et al., 2023), ChatGLM3 (Du et al., 2021), Mistral (Jiang et al., 2023), and Bloom (Workshop et al., 2022). We compare our method with two prevalent training methods: (1) SFT, which refines

Splits / Tasks	Unit	MR	AGNews	AR	MRPC	QQP	MNLI	SST-2	STS-B	ALL
SFT Train	Prompts	2,000	5,000	5,000	1,000	5,000	5,000	5,000	5,000	38,000
RLHF-RM Train	Pairs	16,740	15,993	12,620	9,683	16,956	15,757	16,475	15,751	119,975
RLHF-PPO Train	Prompts	5,000	5,000	4,426	2,668	5,000	5,000	5,000	1,323	33,417
RLLR-RM Train	Pairs	12,339	12,349	12,339	8,877	12,318	12,323	12,338	12,317	95,200
RLLR-PPO Train	Prompts	5,000	5,000	4,426	2,668	5,000	5,000	5,000	1,323	33,417
Test	Prompts	1,000	1,000	1,000	408	1,000	2,000	872	1,000	8,280

Table 1: The number of examples used in experiments.

LLMs through optimization against a conditional language modeling objective on supervised data; (2) RLHF, which involves training a reward model on preference data and subsequently employing this model to guide RL-based fine-tuning. For RLHF, we utilize GPT-4 for the preference annotation within our experiments. Detailed procedural information can be found in Appendix D.

Training. To streamline the experimental complexity, we fine-tune the models on a multi-task dataset, rather than on datasets for individual tasks. To address the task imbalance issue, we construct the training set at a maximum of 5,000 samples per task. The surplus examples are used as unsupervised data for PPO in RLHF and RLLR. This approach mirrors real-world scenarios where unsupervised data is abundant, but supervised data is scarce. We also construct a multi-task test set comprising up to 1,000 examples from each task to enhance experimental efficiency without compromising validity. The details of the examples are listed in Table 1. In all experiments, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2021) fine-tuning, as opposed to full-parameter tuning, achieving up to an 80% reduction in GPU memory requirements. Within the RL-based approaches, the policy, reward, and value models are equipped with their own set of LoRA parameters. Details for the hyperparameters across the various experiments are listed in Appendix A.

4.2 Main Results

Our main experiment results are shown in Table 2. Additional results for models of various sizes are available in Appendix F. SFT *w. rat.* and SFT denote models fine-tuned on supervised data with and without rationales, respectively. RLHF denotes models fine-tuned with the standard RLHF procedure, which predominantly utilizes rationale-sensitive pairs. RLLR denotes models fine-tuned using our proposed method, with a reward model trained on label-sensitive pairs. $\text{RLLR}_{\text{MIXED}}$ further integrates reward models trained on both label-sensitive and rationale-sensitive pairs. The policy model is initialized from the SFT *w. rat.* model in both RLHF, RLLR, and $\text{RLLR}_{\text{MIXED}}$ settings.

Comprehensive evaluations across five foundational models and eight NLU tasks reveal that our RLLR method consistently surpasses the SFT baseline by an average margin of 1.54%, and the RLHF baseline by an average of 0.69%. The maximum average improvement over RLHF was achieved on Mistral 7B, reaching 1.02%. The enhancement observed in ChatGLM3 6B, while modest, is still quantifiable at an increase of 0.38%. RLLR and $\text{RLLR}_{\text{MIXED}}$ also achieve the best results on most individual tasks, except Baichuan2 on AGNews. However, integrating RLLR with other models consistently yields a performance enhancement on AGNews, most notably, exceeding the SFT baseline by 2.9% with Bloom-7B. This substantial improvement robustly validates the efficacy of the proposed method.

The integration of rationales brings improvement over the vanilla SFT by an average margin of 0.79%, demonstrating the benefit of rationales. Despite this improvement, the performance of SFT *w. rat.* still lags behind that of RLLR, suggesting that simply integrating the SFT method with rationales is insufficient. Moreover, the RLHF baseline mirrors the performance of SFT *w. rat.*, with no additional gains, which corroborates the presence of an *objective mismatch* issue.

In Section 4.3, we further analyze the influence of various mechanisms, including the utilization of rationales, reward modeling objectives, and incorporation of multiple rewards in RL fine-tuning. The $\text{RLLR}_{\text{MIXED}}$ method achieves on-par performance with RLLR, surpassing SFT by an average of 1.45%, and RLHF by an average margin of 0.60%. However, we further examine its impact on the quality of rationales, extending our analysis beyond label accuracy.

Methods / Dataset		MR	AGNews	AR	MRPC	QQP	MNLI(m/mm)	SST-2	STS-B	AVG.
LLaMA2 7B	SFT	91.00	92.20	69.40	82.11	85.50	83.50/85.10	96.22	89.24	86.03
	SFT *w. rat.*	91.90	92.50	68.70	83.58	87.90	83.50/85.00	96.56	91.83	86.74
	RLHF	91.90	93.00	68.50	83.82	87.60	83.60/85.00	96.44	92.02	86.79
	RLLR	92.40	93.40	**70.10**	83.82	**88.20**	**85.10/85.90**	**96.79**	**92.31**	**87.47**
	$\text{RLLR}_{\text{MIXED}}$	**92.60**	**93.50**	69.60	**84.07**	88.00	**85.10/85.90**	**96.79**	92.07	87.40
ChatGLM3 6B	SFT	89.00	93.00	68.80	81.37	85.00	81.80/83.90	95.30	89.79	85.33
	SFT *w. rat.*	91.30	93.10	68.20	81.86	84.90	82.80/84.20	**95.87**	90.10	85.51
	RLHF	91.10	93.10	68.90	82.35	85.00	82.80/84.30	**95.87**	90.14	85.64
	RLLR	**91.40**	**93.40**	**69.10**	82.35	85.50	**83.60/84.60**	**95.87**	**91.12**	**86.02**
	$\text{RLLR}_{\text{MIXED}}$	**91.40**	**93.40**	**69.10**	**82.36**	**85.70**	83.50/**84.60**	**95.87**	90.91	85.69
Mistral 7B	SFT	92.10	92.50	70.40	83.58	85.90	84.70/87.50	95.18	91.17	87.00
	SFT *w. rat.*	92.00	92.70	69.40	86.52	86.10	85.40/87.60	96.33	92.06	87.29
	RLHF	92.10	92.20	68.70	85.29	88.30	85.40/87.80	96.22	91.83	87.26
	RLLR	**93.30**	**93.10**	**70.60**	**87.01**	88.30	86.60/**88.90**	**96.90**	**92.32**	**88.27**
	$\text{RLLR}_{\text{MIXED}}$	92.40	92.70	70.30	86.76	**88.70**	**86.80**/88.80	96.67	92.23	88.10
Baichuan2 7B	SFT	90.80	**93.40**	69.90	81.86	84.90	82.90/84.10	95.64	89.09	85.84
	SFT *w. rat.*	90.70	92.50	69.10	82.35	87.00	84.80/85.00	95.99	91.58	86.19
	RLHF	91.20	92.90	68.30	83.09	86.50	84.50/85.30	96.22	91.50	86.25
	RLLR	**91.30**	93.00	70.40	82.84	87.40	**85.70/85.80**	96.33	**91.94**	86.82
	$\text{RLLR}_{\text{MIXED}}$	91.20	93.00	**70.50**	**83.58**	**87.50**	85.50/85.70	**96.44**	91.81	**86.88**
Bloom 7B	SFT	89.20	89.80	69.30	76.96	83.40	75.80/78.50	94.38	87.88	82.80
	SFT *w. rat.*	89.50	91.80	69.80	82.60	83.60	76.50/80.70	94.61	88.58	83.92
	RLHF	89.70	**92.70**	69.40	82.11	84.00	77.00/80.00	94.61	88.96	84.01
	RLLR	**90.10**	**92.70**	**70.90**	84.07	**84.30**	**77.90/81.30**	**95.53**	**89.04**	**84.83**
	$\text{RLLR}_{\text{MIXED}}$	89.50	92.50	70.40	**84.31**	**84.30**	77.80/80.70	94.61	88.95	84.52

Table 2: Experiment results for our methods and baselines, over a range of foundation models and NLU tasks. The abbreviation "SFT *w. rat.*" stands for SFT with rationale.

4.3 Analysis

Utilization of rationales. Incorporating rationales into the SFT stage achieves improvement across 78% of our task-model pairings (35 out of 45), aligning with the advancements reported by (Hsieh et al., 2023; Kim et al., 2023; Fu et al., 2023). Nonetheless, a comparative analysis between SFT with RLLR indicates that the mere addition of rationales to SFT is insufficient. SFT, categorized under Behavior Cloning within the Imitation Learning framework, is prone to suffering from compounding errors (Ross et al., 2011). Theoretically, the minimum expected error for a policy derived through Behavior Cloning grows quadratically with the length of the trajectories. Introducing rationales under this method paradoxically extends trajectory lengths, exacerbating the issue. In contrast, RLLR, rooted in Inverse Reinforcement Learning, effectively reduces compounding errors by optimizing across entire trajectories rather than individual actions (Ho and Ermon, 2016; Swamy et al., 2023), thereby enhancing the effectiveness of rationales.

Reward model performance. To elucidate the superiority of RLLR over RLHF, we scrutinized the efficacy of reward models trained in both methods. The models are evaluated on a hold-out label-sensitive dataset, comprising pairs of correct and incorrect answers with respect to the gold label. This evaluation framework is designed to assess the models' proficiency in differentiating between rationales that lead to either the correct or incorrect labels. As indicated in Table 3, reward models developed under RLLR demonstrate an average accuracy of 90%, outperforming those from RLHF by a margin of 10%, which stand at an average accuracy of 80%. Detailed results on each individual task are presented in Appendix G. These findings align with the main results and underscore the detrimental impact of *objective mismatch* issue within RLHF. Conversely, RLLR is immune to such discrepancies, as its objectives are congruent with the evaluative criteria used to discern between correct and incorrect rationales, thereby yielding superior outcomes across a spectrum of tasks and models.

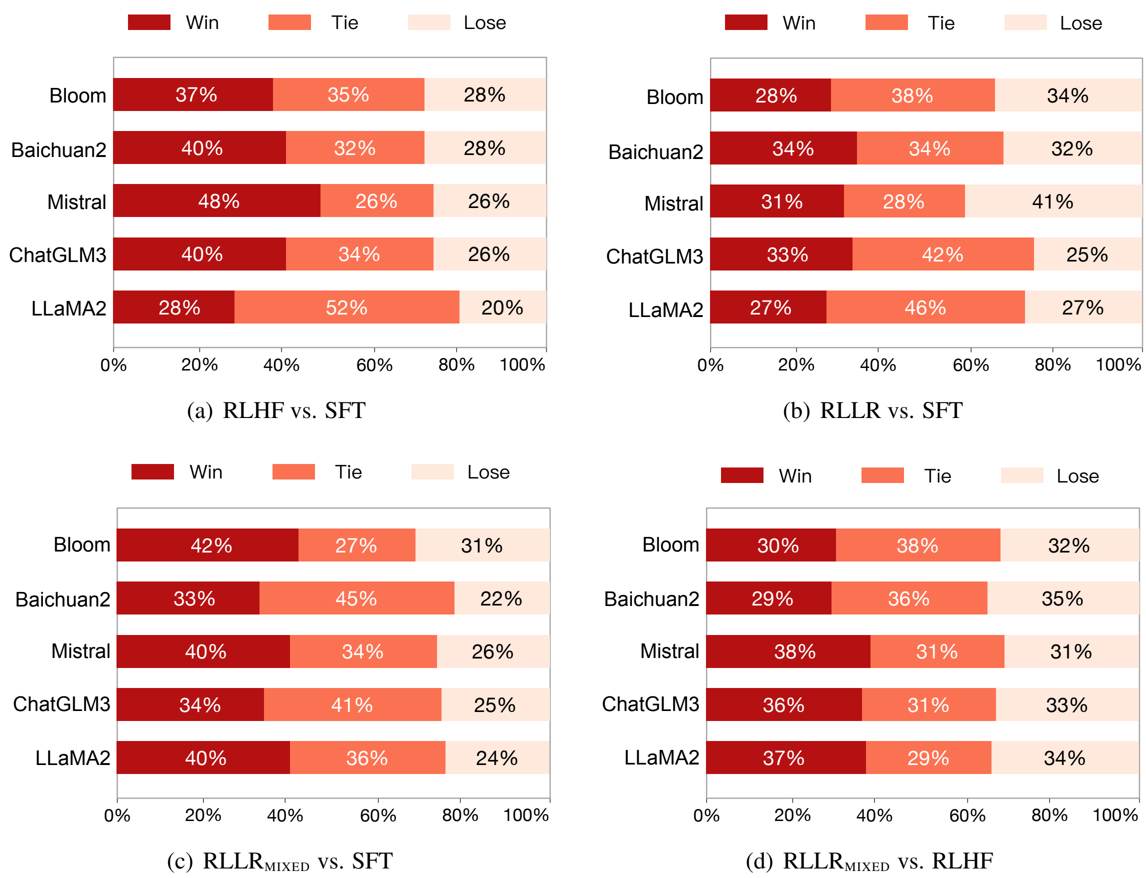

Figure 4: Evaluation of rationale quality judged by GPT-4, compared to the SFT and RLHF methods.

Models / Training Set	RLHF reward	RLLR reward
LLaMA2 7B	80.92	91.66
ChatGLM3 6B	75.00	90.20
Mistral 7B	80.78	91.39
Baichuan2 7B	81.46	90.18
Bloom 7B	77.75	88.73

Table 3: Performance of reward models on hold-out label-sensitive pairs.

Quality of generated rationales. Despite the modest enhancement in accuracy, RLHF has significantly advanced the quality of text generation by integrating human preferences during fine-tuning. Beyond the accuracy on NLU tasks, the quality of the generated text is also a key consideration. This is particularly relevant in human-in-the-loop contexts, where the model's output serves as a guide for human operators, necessitating text that is both high-quality and reflective of human values. To assess this quality, we examined a subset of queries from the validation set and appraised the response quality produced by various models with GPT-4. We employed the win rate against SFT *w. rat.* as a metric for evaluation. The evaluation results are shown in Figure 4 and the implementation details are described in Appendix H. The RLHF method outperforms SFT in generating high-quality rationales, evidenced by a 39% win rate, a 26% lose rate, and a 36% tie rate on average. In contrast, the RLLR method demonstrates comparable performance to SFT, with a balanced win or lose ratio of 30% on average. This outcome aligns with expectations, as RLLR did not incorporate any human preference data. The hybrid RLLR$_{\text{MIXED}}$ method, which integrates the reward models of both RLHF and RLLR, achieves a 38% win rate, a 26% lose rate, and a 37% tie rate on average, benchmarked against the SFT baseline. When directly compared to RLHF, RLLR$_{\text{MIXED}}$ also exhibits a comparable efficacy, with a win rate of 34% and a lose rate of 33%. Notably, RLLR$_{\text{MIXED}}$ yields the most favorable outcomes, delivering rationale quality equivalent to RLHF and maintaining label accuracy at the level of RLLR.

Case study. Table 4 presents examples of answers generated by RLHF, RLLR, and RLLR$_{\text{MIXED}}$

Question	Movie review: ... mafia , rap stars and hood rats butt their ugly heads in a regurgitation of cinematic violence that gives brutal birth to an unlikely , but likable , hero . ' ***What sentiment dose the movie review express?***
RLHF	The movie review expresses a decidedly negative sentiment. The language used is laced with derogatory terms and a tone of disdain, which clearly indicates the reviewer's unfavorable opinion of the film... **Furthermore, the use of "likable, hero" in a seemingly sarcastic or begrudging manner does not redeem the overall negative impression**... ***The answer is: negative***
RLLR	The movie review expresses a positive sentiment, **as it highlights an "unlikely, but likable, hero" that emerges from the chaotic and violent setting**... The use of the word "likable" is a clear indicator of the reviewer's approval, suggesting that the character's journey and the film's outcome are enjoyable and endearing... ***The answer is: positive***
Question	***Analyze the semantic similarity between two sentences and predict a similarity score from 0 to 5.*** Sentence 1: Someone is greating a carrot., Sentence 2: A woman is grating an orange food.
RLLR	**Firstly, the subjects** of the sentences are different... **Secondly, the actions** described in the sentences are not the same... **Thirdly, the objects** of the sentences are distinct... **Lastly, the contexts** implied by the sentences are different... Given these points of divergence, a similarity score of 1.000 accurately captures the minimal semantic overlap between the two sentences... ***The answer is: 1.000***
$\text{RLLR}_{\text{MIXED}}$	**1. **Subjects****: In the first sentence... **2. **Actions****: The actions described in both sentences are related but not identical... **3. **Objects****: The objects of the actions are different... **4. **Context****: The context of the actions is also different... **5. **Semantic Field****: Both sentences fall within the semantic field of food-related activities, but... Given these points, a score of 1.000 accurately captures the low level of semantic similarity between the two sentences... ***The answer is: 1.000***

Table 4: Examples generated by RLHF, RLLR, and $\text{RLLR}_{\text{MIXED}}$ methods.

methods. In the interest of brevity, we have excluded extraneous content from the answers, preserving only the essential context and comparative elements. We emphasize the key distinctions within the examples using blue and red for visual clarity. In the first example, which is sampled from the Movie Review dataset, the reviewer first describes the negative aspects of the movie, but at the end, a turning point is made by proposing "give birth to an unlikely but likable hero" to express positive sentiments. The RLHF model fails to recognize a positive sentiment shift, despite the reviewer's concluding praise, leading to an incorrect negative assessment. Conversely, the RLLR model accurately detects the sentiment change, resulting in a correct evaluation. The second example from the STS-B dataset reveals that, despite RLLR and $\text{RLLR}_{\text{MIXED}}$ methods generating equivalent similarity scores for identical sentence pairs, the $\text{RLLR}_{\text{MIXED}}$ approach augments the rationale's comprehensiveness by incorporating an additional "Semantic Field" component. Moreover, the $\text{RLLR}_{\text{MIXED}}$ output employs Markdown formatting to improve readability. These results indicate that $\text{RLLR}_{\text{MIXED}}$ can significantly enhance the quality of the rationale compared to the standard RLLR method.

5 Conclusion

In this paper, we introduce a novel RLLR method to amplify the performance of LLMs for NLU. By training the reward model on label-sensitive pairs, which are constructed by generating rationales for the incorrect labels, we mitigate the *objective mismatch* issue in RLHF, leading to improved performance in NLU tasks. Extensive results on 5 foundation models and 8 NLU tasks demonstrate that RLLR consistently surpasses the SFT baseline by a margin of 1.54%, and the RLHF baseline by 0.69%. By additionally incorporating the label-sensitive and rationale-sensitive rewards, our enhanced $\text{RLLR}_{\text{MIXED}}$ method not only maintains the label accuracy comparable to RLLR but also achieves rationale quality on par with RLHF. We also present an in-depth analysis of the RLLR framework, examining the utilization of rationales, reward modeling objectives, and incorporation of multiple rewards during the RL fine-tuning stage. The results and analysis substantiate the effectiveness of our methods in NLU tasks.

6 Limitations

Due to cost considerations, there are some deficiencies in our work, which we have listed here for future reference. Firstly, integrating rationales into model responses increases computing power requirements and generation time as a trade-off for enhanced accuracy and interpretability. Secondly, we utilize GPT-4 as a proxy of humans to generate rationales, annotate preferences, and evaluate the quality of rationales. Despite the success made by advanced AI models like GPT-4 in supplanting manual annotation, we believe that experiments with authentic human annotation and evaluation remain essential. Finally, the compatibility of RLLR with RL-free methods such as DPO, PRO, and RRHF remains unexplored. We leave these limitations for future work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

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

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

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*.

Giovanni Grano, Andrea Di Sorbo, Francesco Mercaldo, Corrado A Visaggio, Gerardo Canfora, and Sebastiano Panichella. 2017. Software applications user reviews. In *Zurich Open Repository and Archive:dataset*.

Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*.

Nathan Lambert and Roberto Calandra. 2023. The alignment ceiling: Objective mismatch in reinforcement learning from human feedback. *arXiv preprint arXiv:2311.00168*.

Nathan Lambert, Thomas Krendl Gilbert, and Tom Zick. 2023. The history and risks of reinforcement learning and human feedback. *arXiv e-prints*, pages arXiv–2310.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al.

2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Gokul Swamy, David Wu, Sanjiban Choudhury, Drew Bagnell, and Steven Wu. 2023. Inverse reinforcement learning without reinforcement learning. In *International Conference on Machine Learning*, pages 33299–33318. PMLR.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *arXiv preprint arXiv:2306.01693*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Hyperparameters

The range of hyperparameters for each method is listed in Table 5. The range is the same across all foundation models.

Stages	Items	Values
SFT	learning rate	1e-5~2e-5
	batch size	128
	epochs	20
SFT *w. rat.*	learning rate	1e-4~1e-3
	batch size	128
	epochs	10
RLHF Reward	learning rate	2e-4
	batch size	64~128
	epochs	10
RLHF PPO	learning rate	2e-6~1e-5
	batch size	16~32
	epochs	1
	ppo minibatch size	16~32
	ppo epochs	1
RLLR Reward	learning rate	1e-4~1e-3
	batch size	64~128
	epochs	1
RLLR PPO	learning rate	2e-6~1e-5
	batch size	16~32
	epochs	1
	ppo minibatch size	16~32
	ppo epochs	1
ALL	lora rank	16
	lora alpha	32
	lr scheduler	cosine
	generation strategy	greedy

Table 5: Range of hyperparameters.

B Prompts for Tasks

For all of the tasks, we use the output template `{{label}}` for SFT w/o rationale and `{{rationale}}\n\nThe answer is: {{label}}` for SFT w. rationale, RLHF and RLLR. The name in double curly brackets represents a variable and should be replaced with its value. Input templates and possible labels for each task are listed below.

B.1 Movie Reviews

Input template:
`{{text}} What sentiment does the writer express for the movie?`
Possible labels:
`negative, positive`

B.2 AGNews

Input template:
`What label best describes this news article? \n {{text}}`
Possible labels:
`World politics, Sports, Business, Science and technology`

B.3 MNLI

Input template:
`Given a premise and a hypothesis, predict the relationship between them. Choose one of the following labels: entailment, contradiction, or neutral. Premise:{{sentence1}}, Hypothesis:{{sentence2}}`
Possible labels:
`entailment, contradiction, neutral`

B.4 QQP

Input template:
`I received the questions "{{sentence1}}" and "{{sentence2}}". Are they duplicates?`
Possible labels:
`no, yes`

B.5 SST-2

Input template:
`Movie review: {{text}} What sentiment dose the movie review express?`
Possible labels:
`negative, positive`

B.6 STS-B

Input template:
`Analyze the semantic similarity between two sentences and predict a similarity score from 0 to 5. Sentence 1: {{sentence1}}, Sentence 2: {{sentence2}}`
Possible labels:
Float number in range [0.0, 5.0].

B.7 MRPC

Input template:
`Does the sentence \n {{sentence1}} \n paraphrase (that is, mean the same thing as) this sentence? \n {{sentence2}}`
Possible labels:
`no, yes`

B.8 AppReviews

Input template:

Question	yet another weepy southern bore-athon . ***What sentiment does the writer express for the movie?***
Label	***negative***
Rationale	The sentiment expressed by the writer for the movie is decidedly negative. The term "weepy" suggests that the movie is overly sentimental or melodramatic in a way that the writer finds unappealing. Additionally, "bore-athon" is a pejorative play on the word "marathon," indicating that the writer perceives the movie as boring and tedious, and possibly that it drags on for too long. The use of the word "another" implies a sense of frustration or fatigue with what the writer views as a recurrent theme or genre, in this case, a "southern" one, which the writer apparently does not find engaging or innovative. The overall tone is dismissive and conveys a strong negative opinion about the film.
Question	***What label best describes this news article? \n*** LABOUR AT BRIGHTON: SORRY I JUST CAN #39;T SAY SORRY FOR IRAQ TONY Blair refused to say sorry for the Iraq war yesterday, after a last-minute change of mind over a conference apology. In his Brighton address, the Prime Minister
Label	***World Politics***
Rationale	The label "World politics" aptly describes the news article as it directly involves a statement by Tony Blair, who was the Prime Minister of the United Kingdom, concerning a matter of international significance—the Iraq war. The reference to Blair's address at a conference in Brighton and his stance on a major foreign policy issue that has had global repercussions places the article firmly within the realm of international relations and governmental affairs. The content is inherently political, dealing with decisions and actions that have shaped world events, and it pertains to the political landscape beyond just a national level, thereby making "World politics" the most fitting label for this news piece.
Question	***Does the sentence \n*** As they wound through police barricades to the funeral home , many chanted " Celia , Celia " and sang snippets of her songs . ***\n paraphrase (that is, mean the same thing as) this sentence? \n*** As they wound through police barricades to the funeral home , many chanted " Celia , Celia . "
Label	***no***
Rationale	No, the two sentences do not convey the exact same meaning. While both sentences describe a procession moving through police barricades to a funeral home with people chanting "Celia, Celia," the first sentence includes an additional action that the second sentence omits. Specifically, the first sentence mentions that many in the crowd also sang snippets of her songs. This additional detail about singing introduces a separate action that captures more about the atmosphere and the tribute being paid to Celia. It is not merely a repetition of her name but an engagement with her musical legacy. Without this element, the second sentence fails to capture the full extent of the homage being paid, thus it is not a complete paraphrase of the first.

Table 6: Examples of generated rationales.

```
On a scale of 1-5 (with 1 being least favorable and 5 being most favorable), how would you rate this review? "{{text}}"
```

Possible labels:

```
1, 2, 3, 4, 5
```

C Rationale Collection

We utilize the following prompt templates to request GPT-4 for rationales. Sometimes GPT-4 refuses to give a rationale conditioned the provided label, and we train a simple classifier to filter out these responses. Examples of generated rationales are shown in Table 6.

```
{{question}} \n\n Please give a rationale for the answer "{{label}}" in a confident tone (regardless of the true answer):
```

D Preference Collection

We sample 5 responses from the SFT model for each example and ask GPT-4 to rank the responses. The prompt template for requesting GPT-4 is as follows:

```
Given the following question and answers, please rank the answers according to your preference, considering accuracy, coherence, logicality, factuality,
```

Methods / Dataset			MR	AGNews	AR	MRPC	QQP	MNLI(m/mm)	SST-2	STS-B	AVG.
LLaMA2	7B	SFT	91.00	92.20	69.40	82.11	85.50	83.50/85.10	96.22	89.24	86.03
		SFT *w. rat.*	91.90	92.50	68.70	83.58	87.90	83.50/85.00	96.56	91.83	86.74
		RLHF	91.90	93.00	68.50	83.82	87.60	83.60/85.00	96.44	92.02	86.79
		RLLR	92.40	93.40	**70.10**	83.82	**88.20**	**85.10/85.90**	**96.79**	**92.31**	**87.47**
		RLLR$_{\text{MIXED}}$	**92.60**	**93.50**	69.60	**84.07**	88.00	**85.10/85.90**	**96.79**	92.07	87.40
	13B	SFT	92.00	92.20	69.00	81.62	88.00	83.10/85.20	96.33	90.12	86.40
		SFT *w. rat.*	92.20	92.40	68.90	83.82	88.40	85.40/87.10	**96.79**	91.78	87.42
		RLHF	92.20	92.70	68.90	**85.78**	88.70	85.60/87.20	96.67	91.49	87.69
		RLLR	**92.60**	**93.00**	69.50	85.54	**88.80**	**86.10/87.80**	**96.79**	**92.23**	**88.04**
		RLLR$_{\text{MIXED}}$	92.40	92.90	**69.70**	85.54	**88.80**	**86.10**/87.50	**96.79**	**92.23**	88.00
Bloom	3B	SFT	88.40	90.20	67.90	75.25	81.30	73.30/74.70	93.46	86.66	81.24
		SFT *w. rat.*	88.70	92.00	68.50	80.15	81.90	73.40/75.10	93.23	86.58	82.17
		RLHF	88.60	92.00	68.60	79.66	82.20	74.20/75.60	93.00	86.23	82.23
		RLLR	**89.80**	**92.30**	69.20	**80.64**	**82.60**	74.60/**76.70**	93.46	**87.58**	**82.99**
		RLLR$_{\text{MIXED}}$	89.40	92.00	**69.30**	**80.64**	82.40	**74.70**/76.40	**93.58**	87.17	82.84
	7B	SFT	89.20	89.80	69.30	76.96	83.40	75.80/78.50	94.38	87.88	82.80
		SFT *w. rat.*	89.50	91.80	69.80	82.60	83.60	76.50/80.70	94.61	88.58	83.92
		RLHF	89.70	**92.70**	69.40	82.11	84.00	77.00/80.00	94.61	88.96	84.01
		RLLR	**90.10**	**92.70**	**70.90**	84.07	**84.30**	**77.90/81.30**	**95.53**	**89.04**	**84.83**
		RLLR$_{\text{MIXED}}$	89.50	92.50	70.40	**84.31**	**84.30**	77.80/80.70	94.61	88.95	84.52

Table 7: Results of varying sized models.

relevance, and information completeness. \n\n [Question] {question} \n\n [Answer 1] {answer 1} \n\n [Answer 2] {answer 2} \n\n [Answer 3] {answer 3} \n\n [Answer 4] {answer 4} \n\n [Answer 5] {answer 5} \n\n Please give your rationale first, and then give the ranking. Output format: "{rationale} \n\n Ranking: {e.g. {{ranking_example}}}"
The variable {{ranking_example}} is generated by shuffling the list [1, 2, 3, 4, 5] and concatenating them with ">" or "=", e.g. 5>3>2>4>1 or 2>1=5>3=4. We generate a different example for every GPT-4 request to avoid bias.

E Construction of Label-Sensitive Pairs

In tasks involving categorical labels, an incorrect label is randomly chosen from the full label set excluding the correct label, to create a label-sensitive pair. For the AppReviews and STS-B tasks, which use a rating scale from 0 to 5, incorrect labels are generated by adding 3 to the correct label and then incorporating a random value from the range [-1, 1]. For instance, given a correct STS-B label of 2.8, a random increment of 0.3 is selected, resulting in an initial incorrect label of 2.8+3+0.3=6.1. This exceeds the maximum rating, so we adjust by subtracting 5, yielding a final incorrect label of 1.1. In the case of the AppReviews task, this label is subsequently rounded to an integer.

F Results of Varying Sized Models

To substantiate the scalability of our method across models of varying sizes, we also conduct a series of experiments using LLaMA2-13B and Bloom-3B models. The results are presented in conjunction with those of the 7B models to facilitate direct comparison. As shown in Table 7, LLaMA2-13B and Bloom-3B achieved consistent results with the 7B model, achieving average improvements of 1.64% and 1.75% over the SFT baseline, respectively. This consistency across disparate model sizes strongly supports the scalability of our proposed RLLR method.

G Reward Model Performance

Table 8 presents the performance of reward models trained with RLHF and RLLR on eight individual NLU tasks. Reward models employing RLLR methods demonstrated an overall accuracy of approximately 90%, surpassing those trained with RLHF by a significant margin of 10 percentage points, with the latter achieving an accuracy of 80%. The gap between RLLR and RLHF on STS-B and AppReviews tasks is most significant, exceeding 35% and 20% respectively. The gap on AGNews,

Tasks / Models	LLaMA2 7B		ChatGLM3 6B		Mistral 7B		Baichuan2 7B		Bloom 7B	
	RLHF	RLLR	RLHF	RLLR	RLHF	RLLR	RLHF	RLLR	RLHF	RLLR
MR	90.13	92.07	80.26	90.94	90.94	90.78	89.32	91.59	91.10	88.51
AGNews	89.67	96.31	85.61	94.10	87.45	96.68	89.67	94.46	88.56	94.10
AR	69.91	92.76	74.91	91.26	68.91	92.51	67.79	90.26	65.29	90.89
MRPC	84.03	86.58	71.25	87.22	76.04	86.58	77.96	83.07	78.91	82.43
QQP	79.82	86.63	70.18	86.63	78.92	87.53	80.46	88.17	76.74	83.80
MNLI	86.65	90.75	83.00	89.04	87.40	91.13	88.59	89.56	84.41	87.47
SST-2	92.24	94.68	82.93	93.13	93.13	94.01	93.13	92.90	89.36	92.24
STS-B	58.68	97.07	46.80	93.05	62.34	94.52	64.90	92.50	50.46	93.97
Overall	80.92	91.66	75.00	90.20	80.78	91.39	81.46	90.18	77.75	88.73

Table 8: Performance of reward models on hold-out label-sensitive pairs. Results across five different foundation models are presented.

MRPC, and QQP tasks also exceeds 5%, indicating that RLHF suffers from *objective mismatch* issue on these tasks.

H Evaluation of Generation Quality

We utilize GPT-4 as a proxy for human evaluation. First, we sample a set of questions and corresponding answers generated by two methods. To mitigate positional bias, we then randomize the order of the answers within each pair. The question along with two answers is subsequently formatted according to the predefined GPT4 input template:

```
Given the following question and two candidate answers, please choose which one is better, considering accuracy, coherence, logicality, factuality, relevance, and information completeness: \n\n [Question] {question} \n\n [Answer 1] {answer 1} \n\n [Answer 2] {answer 2} \n\n Please response with "Answer 1 is better" or "Answer 2 is better" or "Equal" first, and then give your rationale.
```