

🕒 ALARM: Align Language Models via Hierarchical Rewards Modeling

Yuhang Lai[♣] Siyuan Wang[♣] Shujun Liu[♣]
Xuanjing Huang[◇] Zhongyu Wei^{♣△†}

[♣]Institute of Science and Technology for Brain-inspired Intelligence, Fudan University

[♣]School of Data Science, Fudan University

[◇]School of Computer Science, Fudan University

[△]Research Institute of Intelligent and Complex Systems, Fudan University

yhlai23@m.fudan.edu.cn, {wangsy18,xjhuang,zywei}@fudan.edu.cn

Abstract

We introduce ALARM, the first framework modeling hierarchical rewards in reinforcement learning from human feedback (RLHF), which is designed to enhance the alignment of large language models (LLMs) with human preferences. The framework addresses the limitations of current alignment approaches, which often struggle with the inconsistency and sparsity of human supervision signals, by integrating holistic rewards with aspect-specific rewards. This integration enables more precise and consistent guidance of language models towards desired outcomes, particularly in complex and open text generation tasks. By employing a methodology that filters and combines multiple rewards based on their consistency, the framework provides a reliable mechanism for improving model alignment. We validate our approach through applications in long-form question answering and machine translation tasks, employing gpt-3.5-turbo for pairwise comparisons, and demonstrate improvements over existing baselines. Our work underscores the effectiveness of hierarchical rewards modeling in refining LLM training processes for better human preference alignment. We release our code at <https://ALaRM-fdu.github.io>.

1 Introduction

Current LLM-assisted AI systems have shown remarkable performance in a wide range of tasks (Brown et al., 2020; Chen et al., 2021; Touvron et al., 2023; Wang et al., 2024), and benefit from different forms of human supervision signals (Wei et al., 2022; Stiennon et al., 2020). While supervised learning relies on human-written demonstrations to unlock the emergent abilities gained from pretraining on huge text and code corpora, RLHF utilizes generation comparisons labeled by humans to further fine-tune the LLMs for better alignment

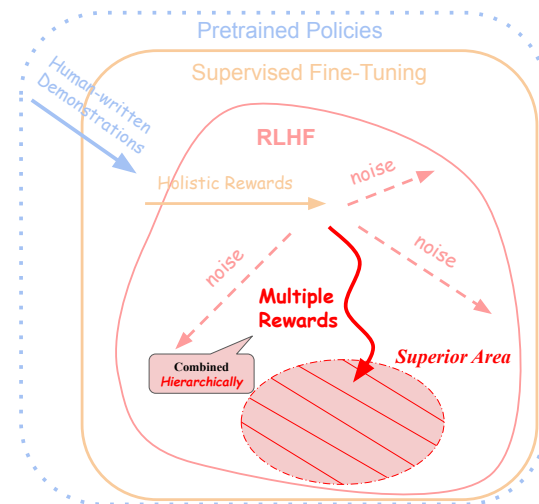


Figure 1: Illustration of our key ideas. The pretrained policies are first supervised fine-tuned on human-written demonstrations and then trained through RLHF given a holistic reward learned from human comparisons. The shadowed "superior area" better aligns with human preference, which is hard to reach for solely a noisy holistic reward. We propose to utilize multiple rewards hierarchically for more accurate and consistent supervision signals and thus guide the policies into the superior area.

with human expectations, which has been demonstrated to be able to reduce undesired model generations like harmful contents (Bai et al., 2022) or hallucinations (Ouyang et al., 2022).

However, human oversight capabilities are finite. As recent LLMs are capable of doing more complex work and even surpass human performance in some areas, it becomes more difficult to write good enough demonstrations even for human experts (Lai et al., 2023). Comparisons between several model generations can be intuitively easier to get from a crowd-sourcing platform though, previous research (Krishna et al., 2023; Wu et al., 2023) has revealed that human annotations can be inconsistent and unreliable in evaluations between two or more model outputs for complex tasks like long text generations, producing unstable rewards in RLHF.

[†]Corresponding author

While stable rewards are the key to successful reinforcement learning, the sparsity nature of current holistic rewards further stresses this challenge (Cao et al., 2024). Moreover, various scenarios hidden behind the tasks can differentiate preference standards (Jang et al., 2023), thus degrading the annotation consistency and value alignment on downstream applications, e.g., concise vs. comprehensive summary in text summarization, and literary vs. technical style in machine translation. Then we ask the question, how to get reliable and scalable supervision signals within limited human oversight capabilities?

To take an initial step towards addressing this issue, we introduce a new framework ALARM hierarchically modeling both holistic and aspect-specific rewards, which is motivated by 1) fine-grained RLHF (Wu et al., 2023) that categorizes different error types for more accurate and easier annotation, 2) task decomposition in hierarchical reinforcement learning (Pateria et al., 2021) that helps to overcome sparse rewards. At the core of our framework is to seek stronger supervision signals: As shown in Figure 1, solely using the holistic reward can make it difficult to reach the shadowed "superior area" which represents better alignment with human preference. Thus we employ multiple rewards combined in a hierarchical way to stabilize the optimization direction for more accurate and consistent guidance into the superior area. Firstly, we list several aspect-specific rewards corresponding to the task and perform the selection by their inconsistency with the holistic reward in pairwise comparisons. Later in the RLHF training process, the chosen rewards are combined with the holistic reward as a whole once the sampled generation receives a high holistic reward, which is above a certain threshold value. The aspect-specific rewards can either come from reward models trained on comparison datasets annotated along a specific dimension (e.g., honesty) or simply be toolkit-calculated metrics (e.g., token count), with an arbitrary density at either the token level or the sequence level. In addition, we proactively transform the aspect-specific rewards to guarantee that their cumulative values are positive, thereby motivating the policy to surpass the threshold for higher returns, which enriches the effectiveness of the hierarchical structure.

We apply our framework to two text generation tasks: long-form question answering (QA) and machine translation (MT). Long-form QA represents

difficult annotations for complex tasks, and MT represents more flexible and various preference standards behind the tasks. For each of them, we employ gpt-3.5-turbo as the evaluator for pairwise comparisons. We empirically demonstrate that our framework outperforms the compared baselines. The ablation studies and corresponding analyses further show that our framework effectively provides stronger supervision signals toward human preference in both scenarios.

Collectively, we highlight our contributions as follows: 1) To our knowledge, we are the first to propose a framework that hierarchically models both holistic and aspect-specific rewards in RLHF, 2) we investigate how to perform reward selection to mitigate rewards conflicting, 3) we demonstrate the effectiveness of ALARM as pursuing more accurate and consistent supervision signals on two text generation tasks through comprehensive ablation studies and analyses, shedding light on its potential for scalable oversight.

2 Framework

Our framework originates from RL-based text generation (Ryang and Abekawa, 2012; Buck et al., 2018), meanwhile modeling both holistic and aspect-specific rewards hierarchically. We first introduce the widely used RLHF. Then, we discuss why we should do reward selection proactively, and how we do that. After that, we present our method of hierarchical rewards modeling in detail.

2.1 Reinforcement Learning from Human Feedback

In the context of RL-based text generation, we define a language model parameterized by θ as a policy π_θ . Token generation is considered a decision-making process, and the policy completes an episode when the language model generates an EOS token or meets the length limit.

In this way, RLHF aims to optimize the policy π_θ to maximize the reward from a preference model R with Proximal Policy Optimization (PPO) (Schulman et al., 2017), a broadly used reinforcement learning algorithm for human preference alignment. The preference model R is learned on human-annotated comparison datasets to predict a single scalar that correctly ranks two or more model generations in the same comparison batch. The optimization objective is also formulated as follows:

$$\arg \max_{\theta} \mathbb{E}_{\tau \sim \pi_\theta} [R(\tau)] \quad (1)$$

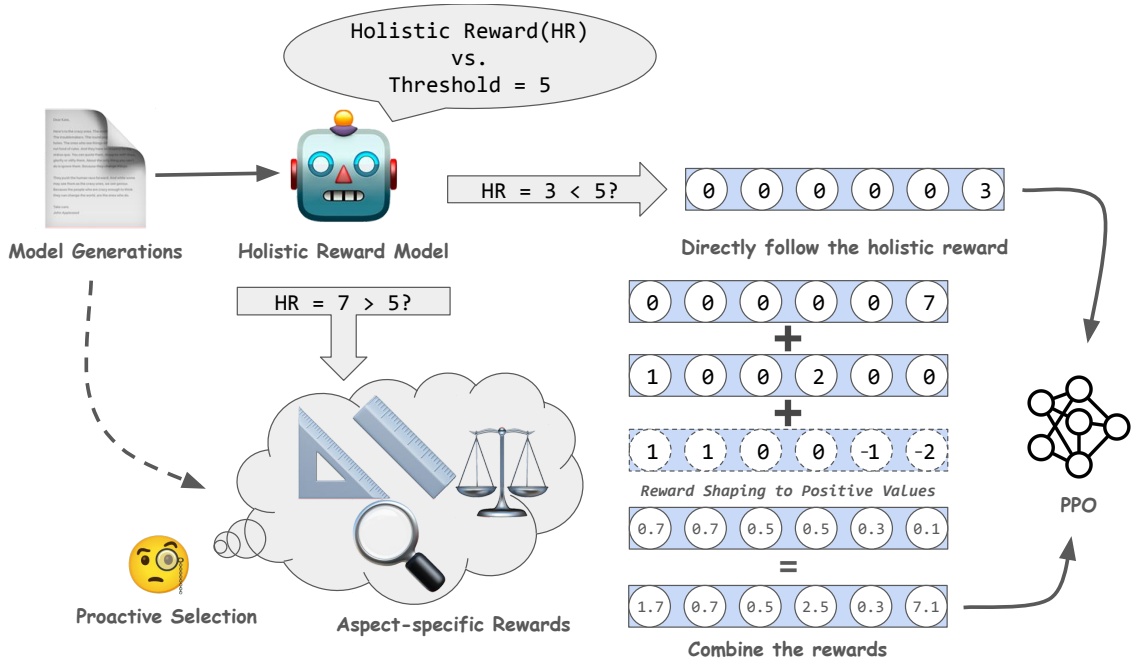


Figure 2: Illustration of our framework. The reward modeling is decomposed into two parts: 1) Directly assign the holistic reward to improve general quality, 2) combine the holistic reward and proactively selected aspect-specific rewards as a whole reward, which is supposed to be more accurate and consistent.

where τ denotes a trajectory produced by the policy π_θ , and $R(\tau)$ is the reward associated with the trajectory τ , as evaluated by the preference model R . Here we see how the reward model R deeply affects RLHF and thus it’s crucial to ensure the accuracy and consistency of its prediction.

2.2 Reward Selection

Evaluation along a specific dimension of model generations instead of the general quality is demonstrated to be less noisy and more accurate for reward modeling (Wu et al., 2023). Therefore, to get more accurate and consistent supervision signals, we first intuitively list several aspect-specific rewards corresponding to a certain task. However, human preferences are intricate. Different decomposed aspects are interconnected and can even conflict with each other. A common way to balance them is the weighted sum method, which assigns a carefully chosen weight for each aspect-specific reward, based on observations of either performance during training or accuracy in pairwise comparisons (Go et al., 2024). Nevertheless, this method still suffers from the over-optimization problem (Moskovitz et al., 2024), where the model loses individual information from every single aspect-specific reward and cannot attribute changes in the composed reward to any of them.

Our key idea, which differs from merely combin-

ing all aspect-specific rewards, is to stabilize the supervision signals. We need a "copilot" for the holistic reward. Thus we aim to resolve this challenge by discarding the conflicting rewards, and we select the rewards that are mostly consistent with the holistic reward. Therefore, we proactively conduct pairwise comparisons between two sets of model generations. These generations come from the same supervised fine-tuned model but are produced using greedy decoding and pure sampling, respectively. Then, we calculate the prediction inconsistency by assessing whether the holistic and aspect-specific rewards have divergent predictions on which answer is better.

2.3 Hierarchical Rewards Modeling

Hierarchical reinforcement learning has advanced significantly in a wide range of decision-making tasks (Yang et al., 2018; Saleh et al., 2020; Wang et al., 2018), decomposing complex and challenging optimization objectives into simpler sub-tasks. Nevertheless, in contrast, existing RLHF works typically employ a plain rewarding strategy that linearly assigns a single holistic reward (Sun et al., 2023) or a fixed combination of aspect-specific rewards (Go et al., 2024), which not only poses sparse rewards in the long-horizon optimization but also overlooks the close relationships between the holistic reward and the aspect-specific rewards.

With these motivations, we propose a novel approach that leverages both holistic and aspect-specific rewards. In this way, we consider the optimization objective that aligns the language models with human preference, targeting the superior area depicted in Figure 1, as a challenging decision-making task. Thus we propose a decomposition of this task into two less complex sub-tasks which ought to be addressed sequentially: 1) Directly follow the holistic reward until the model generation receives a high holistic reward, which indicates the generation is generally good and meets human preference at a relatively high level, 2) optimize the combination of the holistic and aspect-specific rewards, which as a whole provides more accurate and consistent supervision signals towards the superior area. Unlike the plain weighted-sum method, which applies combined rewards throughout the entire training, our approach is more nuanced. We primarily follow the supervision of the holistic reward and gently turn the steering wheel only when it’s insufficient to solely rely on the holistic reward to reach the superior area.

As illustrated in Figure 2, we first follow the regular RLHF process to sample some generations from the policy. Then we employ a preference model that predicts a single scalar as the holistic reward for each of them. Here we utilize UltraRM-13B (Cui et al., 2023) as a zero-shot reward model to predict holistic rewards for all experiments reported in this paper. For each sampled generation, if it receives a holistic reward lower than a certain threshold value, we directly assign it as the final reward. Otherwise, we calculate the proactively selected aspect-specific rewards and combine all the rewards together. To design a more effective hierarchical architecture for the above reward modeling, we ensure that the generations receiving a holistic reward above the threshold obtain higher cumulative rewards than those below this threshold. We achieve this through reward shaping, where we transform aspect-specific rewards into positive values using the sigmoid function.

3 Long-Form Question Answering

Long-form QA is a complicated task that aims to produce elaborate responses covering background information, explanations, or discussions corresponding to the questions. This process is difficult and even humans are not able to write high-quality demonstrations or make accurate and consistent

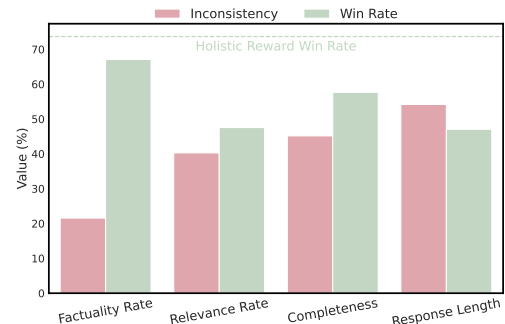


Figure 3: Inconsistency with the holistic reward for listed aspect-specific rewards and win rates of the greedy decoding against the pure sampling in long-form QA.

comparisons reliably. Following we show how we address this issue using ALARM.

3.1 Task Settings

We utilize most of the settings and release from Wu et al. (2023): QA-Feedback dataset, supervised fine-tuned T5-large as the initial policy, and three fine-grained reward models.

Dataset. QA-Feedback dataset is extracted from ASQA (Stelmakh et al., 2022) and then transformed into a task format of reading comprehension, which gives an ambiguous factoid question and a set of related knowledge sources from the Wikipedia corpus, and requires the language model to generate a long-form response. QA-Feedback has a training set of 3,853 examples, a development set of 500, and a test set of 948 in total. And we keep the division of this dataset in the experiment.

Initial Policy. The initial policy model is T5-large (Raffel et al., 2020) initialized with supervised fine-tuning on 1K training examples.

Reward Models. We reuse the three fine-grained reward models designed in Wu et al. (2023), which are called the relevance reward model R_{ϕ_1} , the factuality reward model R_{ϕ_2} , the completeness reward model R_{ϕ_3} , representing three categories of different error types. They all use Longformer-base (Beltagy et al., 2020) as the backbone model and do predictions at different levels of density. R_{ϕ_1} is trained to predict whether the generation contains irrelevance, repetition, or incoherence errors at the sub-sentence level, with a binary classification accuracy of 69.6 and an F1-score of 68.5 on the development set. R_{ϕ_2} learns to detect incorrect or unverified facts for each sentence according to

Methods	Holistic Reward	Factuality Reward	Weighted Sum	ALARM	Avg.
<i>% Win Rates by the Holistic Reward</i>					
Holistic Reward	-	55.86 ± 4.22	50.36 ± 0.95	49.06 ± 0.48	51.76 ± 1.66
Factuality Reward	44.13 ± 4.22	-	45.04 ± 3.73	42.73 ± 3.91	43.97 ± 3.92
Weighted Sum	49.64 ± 0.95	54.95 ± 3.73	-	48.55 ± 1.14	51.05 ± 0.67
ALARM	50.94 ± 0.48	57.27 ± 3.91	51.45 ± 1.14	-	53.22 ± 1.76
<i>% Win Rates by the Factuality Rate</i>					
Holistic Reward	-	48.31 ± 3.49	50.89 ± 1.50	44.42 ± 2.58	47.87 ± 1.48
Factuality Reward	51.68 ± 3.49	-	52.76 ± 4.01	48.70 ± 3.17	51.05 ± 3.49
Weighted Sum	49.11 ± 1.50	47.24 ± 4.01	-	44.65 ± 4.40	47.00 ± 3.00
ALARM	55.58 ± 2.58	51.30 ± 3.17	55.35 ± 4.40	-	54.08 ± 2.15
<i>% Win Rates Evaluated by gpt-3.5-turbo</i>					
Holistic Reward	-	44.11 ± 3.12	44.16 ± 6.28	40.94 ± 1.87	43.07 ± 1.81
Factuality Reward	55.88 ± 3.12	-	53.55 ± 6.09	50.77 ± 4.01	53.40 ± 4.21
Weighted Sum	55.84 ± 6.28	46.45 ± 6.09	-	46.03 ± 5.94	49.44 ± 5.93
ALARM	59.06 ± 1.87	49.23 ± 4.01	53.96 ± 5.94	-	54.08 ± 2.18

Table 1: Evaluation of win rates and corresponding standard error determined by the holistic reward, the factuality rate, and the evaluator gpt-3.5-turbo respectively in long-form QA.

ALARM vs.	Mean HR	Win Rate by HR	Mean FR	Win Rate by FR	Win Rate by gpt-3.5-turbo
Holistic Reward	Tie	Tie	Win (3e-5)	Win (1e-3)	Win (1e-8)
Factuality Reward	Win (3e-15)	Win (1e-11)	Tie	Tie	Tie
Weighted Sum	Win (0.01)	Tie	Win (2e-4)	Win (4e-3)	Win (0.02)

Table 2: Statistical testing to examine the significance of the evaluation results in long-form QA. Here HR represents the holistic reward. FR represents the factuality rate. The orange values in parentheses represent the p-values. See more details in [Appendix B](#).

Methods	Mean HR(↑)	Mean FR(↑)	Length
Holistic Reward	0.608	0.736	87.8
Factuality Reward	0.538	0.752	116.8
Weighted Sum	0.598	0.738	98.5
ALARM	0.617	0.752	97.6

Table 3: Mean rewards on the test set in long-form QA.

given knowledge sources, having an accuracy of 77.8 and an F1-score of 67.5. R_{ϕ_3} is developed to measure the holistic information completeness for the full sequence, with an accuracy of 70.9 in pairwise comparisons.

3.2 Reward Selection

Listing Corresponding Rewards. Following the task settings, we have three aspect-specific reward models. We also consider the response token count as one corresponding reward in this task.

Calculating Inconsistency To proactively filter out appropriate rewards that mostly aid the holistic reward, together as more accurate and consistent

training signals, we conduct pairwise comparisons to check the inconsistency with the holistic reward of those four candidate rewards.

We first employ the initial policy to get two sets of model generations on the training set, using greedy decoding and pure sampling respectively, and two generations from different sets to the same question form a comparison pair. For each aspect-specific reward, we calculate the inconsistency as the percentage of the pairs that the holistic reward prefers one, while the aspect-specific reward prefers another. We disregard all the ties that have the same holistic reward. For the relevance reward and the factuality reward, they cannot make an apple-to-apple comparison due to their sentence or sub-sentence level reward density. Thus we compute the overall correct rate in all the slices of each generation for direct comparison, named factuality rate and relevance rate.

As shown in [Figure 3](#), the factuality rate shows significantly lower inconsistency than the other three rewards, indicating it is more suitable to be the "copilot" of the holistic reward. Also, we com-

pared the win rates of the greedy decoding generations to the pure sampling ones. The win rate is calculated as described in subsection 3.3. Note that the factuality rate has more similar win rates to the holistic reward, which further supports its consistency. Therefore, we select the factuality reward for hierarchical rewards modeling.

3.3 Experimental Setup

Rewards Modeling. We z-normalize the holistic reward in a set of generations D_p which is produced through pure sampling on the training set. The factuality reward is shaped to positive values to ensure the hierarchical structure by the sigmoid function. The threshold value is set at 0.6, which is around the top 30% in D_p . We simply add the rewards up to combine them and only adjust the weight of the holistic reward.

Reinforcement Learning Training. We use the pure sampling strategy in the reinforcement learning process and use greedy decoding for the development set and the test set evaluation. Beginning with the supervised fine-tuned initial policy, we train our model for 2 epochs on the training set. As we set the exploration frequency to 4, the training runs for about 30K episodes. We utilize LoRA (Hu et al., 2022) for the training.

Evaluation. Besides the mean value of the holistic reward and the factuality rate, we evaluate the win rates between different models in pairwise comparisons. Following Jang et al. (2023), the win rate is formulated as:

$$\text{Win Rate} = \frac{\text{Win}}{\text{Win} + \text{Lose}} \quad (2)$$

where all ties are disregarded in the calculation. We also evaluate the win rates in general quality by employing gpt-3.5-turbo as the evaluator for pairwise comparison. To mitigate the positional bias (Wang et al., 2023a) in LLM-as-a-judge, we perform prompting twice with the two generations swapped. We only consider the comparison valid when both promptings prefer the same generation, indicating the evaluator is faithful enough to overcome the positional bias. We use the prompt (Figure 6) from AlpacaEval (Dubois et al., 2023).

Compared Methods. We compare our framework to three methods. **ALARM** represents our proposed hierarchical rewards modeling approach. **Holistic Reward** represents the baseline using the

holistic reward as the only supervision signal. **Factuality Reward** represents solely using the factuality reward without reward shaping to train the policy. **Weighted Sum** is the plain weighted sum method that directly adds the holistic reward and the factuality reward together for training.

3.4 Main Results

In this paper, we conduct all experiments using three different seeds, and the results are averaged across these three independent runs. Table 3 shows the evaluation results on the test set of the mean values of each reward. We can see that ALARM leads to significantly higher holistic reward than other methods, meanwhile reaching the highest factuality rate. As expected, except for ALARM, Holistic Reward gets the highest holistic reward value and Factuality Reward gets the highest factuality rate. Weighted Sum balances these two rewards instead. Table 1 represents the win rates between the four methods. And we can see ALARM holds the best under all three different metrics, which further indicates that ALARM provides a stronger supervision signal than other methods. Furthermore, we conduct statistical testing on the evaluation results. As shown in Table 2, ALARM ties with one and outperforms another, which provides statistical evidence of the leading performance.

4 Machine Translation

Machine translation can be considered a text generation task that involves converting a piece of text from the source to the target language while preserving the original meaning, context, and cultural nuances. This task requires not only a deep understanding of the grammatical structure and vocabulary of both the source and target languages but also an appreciation of their idiomatic expressions

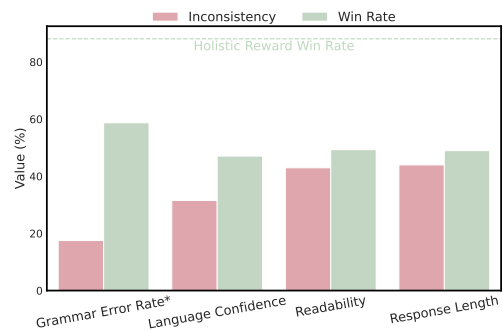


Figure 4: Selection results of inconsistency and win rates in MT. *: The lower grammar error rate wins.

Methods	Holistic Reward	Grammar Reward	Weighted Sum	ALARM	Avg.
<i>% Win Rates by the Holistic Reward</i>					
Holistic Reward	-	51.52 ± 0.77	50.18 ± 0.89	48.30 ± 0.66	50.00 ± 0.76
Grammar Reward	48.47 ± 0.77	-	48.69 ± 0.15	47.07 ± 0.16	48.08 ± 0.32
Weighted Sum	49.82 ± 0.89	51.31 ± 0.15	-	48.26 ± 0.15	49.80 ± 0.25
ALARM	51.70 ± 0.66	52.93 ± 0.16	51.74 ± 0.15	-	52.12 ± 0.26
<i>% Win Rates by the Grammar Error Rate¹</i>					
Holistic Reward	-	51.90 ± 2.83	53.47 ± 3.58	42.20 ± 4.29	49.19 ± 0.40
Grammar Reward	48.10 ± 2.83	-	50.63 ± 5.11	42.47 ± 5.29	47.07 ± 3.82
Weighted Sum	46.54 ± 3.58	49.37 ± 5.11	-	40.63 ± 4.58	45.51 ± 3.88
ALARM	57.80 ± 4.29	57.53 ± 5.29	59.37 ± 4.58	-	58.24 ± 4.46
<i>% Win Rates² Evaluated by gpt-3.5-turbo</i>					
Holistic Reward	-	51.03 ± 0.42	51.27 ± 1.79	48.97 ± 1.70	50.43 ± 1.00
Grammar Reward	48.97 ± 0.42	-	48.98 ± 2.77	47.85 ± 2.09	48.60 ± 0.90
Weighted Sum	48.73 ± 1.79	51.02 ± 2.77	-	47.98 ± 3.05	49.24 ± 2.30
ALARM	51.03 ± 1.70	52.15 ± 2.09	52.02 ± 3.05	-	51.73 ± 2.11

Table 4: Evaluation of win rates in MT. ¹: The lower grammar error rate wins in the calculation. ²: We choose a smaller set of 3K examples randomly selected from the test set to reduce the annotation cost.

ALARM vs.	Mean HR	Win Rate by HR	Mean GER	Win Rate by GER	Win Rate by gpt-3.5-turbo
Holistic Reward	Win (4e - 28)	Win (4e - 19)	Tie	Win (2e - 83)	Win (0.01)
Grammar Reward	Win (3e - 94)	Win (3e - 57)	Lose (4e - 38)	Win (6e - 37)	Win (4e - 16)
Weighted Sum	Win (2e - 29)	Win (3e - 20)	Lose (4e - 3)	Win (1e - 131)	Win (1e - 7)

Table 5: Statistical testing in MT. GER represents the grammar error rate.

Methods	Mean HR(↑)	% Mean GER(↓)	Length
Holistic Reward	0.919	1.203	36.7
Grammar Reward	0.908	1.190	36.5
Weighted Sum	0.918	1.197	36.5
ALARM	0.928	1.203	37.0

Table 6: The means of rewards on the test set in MT.

and cultural references. However, this feature can differentiate people’s preferences and form noise (Marchisio et al., 2019; Savoldi et al., 2021), due to the wide range of application scenarios and user groups, thus posing barriers to accurate and consistent alignment annotation. Below, we show how we use ALARM on this task.

4.1 Task Settings

Dataset. We utilize Europarl (Tiedemann, 2012), a Spanish-English dataset that contains transcripts of European Parliamentary proceedings. We select 100K samples from it and arrange them into a training set of 63K, a development set of 2K, and a test set of 30K in total.

Initial Policy. We initialize the policy with mT5-base (Xue et al., 2021), and then supervised fine-tuned it on the training set for 5 epochs. The initial policy has a BLEU score of 31.57 on the test set.

4.2 Reward Selection

Listing Corresponding Rewards. We first intuitively list three corresponding rewards, named grammar reward, language confidence, and readability. All of them are calculated through python-wrapped toolkits. The grammar reward utilizes LanguageTool, which can detect grammar errors at the word level. We define the grammar reward as assigning negative values to the grammatically incorrect tokens. The language confidence is based on Lingua, which computes the language likelihood in a set of languages by n-gram models as a single scalar. The readability is to measure the general difficulty of reading the text, which can be calculated by Textstat at the full sequence level.

Calculating Consistency. The same as in subsection 3.2, we conduct pairwise comparisons on two sets of generations produced by greedy decoding and pure sampling respectively. We define the

grammar error rate of a generation as the token count divided by the number of grammar errors.

As shown in Figure 4, the grammar reward stands out with lower inconsistency and better win rates than other rewards. Thus we choose the grammar reward for the following experiments.

4.3 Experimental Setup

We follow most setups in subsection 3.3. Evaluation and compared methods are the same.

Rewards Modeling. We apply z-normalization to the holistic reward and set the threshold value at 0.5, which corresponds to the top 30%. Considering the token level density of the grammar reward, we apply reward shaping for every token, including those with a 0 reward, using the sigmoid function to maintain the hierarchy.

Reinforcement Learning Training. With the pure sampling strategy in training, we train our models for around 13K episodes.

4.4 Main Results

Table 6 represents the mean values of separate rewards on the test set, where ALARM has the highest holistic reward and a comparable grammar error rate to the other methods. As shown in Table 4, ALARM continues to outperform the others in the win rates, evaluated by the holistic reward, the grammar error rate, and gpt-3.5-turbo respectively. As for the statistical testing in Table 5, ALARM excels in all except for the mean grammar error rate. The conflict between win rates and mean rewards by the grammar error rate may be due to the variation in reward distribution.

5 Ablation Study

Without Selection. As shown in Table 7, to find out how reward selection affects the performance of ALARM, we conduct extensive experiments that separately apply each reward listed in the initial pool on both tasks. The proactively selected rewards present leading performance evaluated by both the holistic reward and gpt-3.5-turbo, showing the effectiveness of reward selection. We also observe conflicting scores for some rewards from the two evaluators. We attribute this to the biases and flaws in the holistic reward, such as constantly overlooking or overvaluing certain aspects (Zheng et al., 2023). For example, we consider the relevance reward complements the biases and

Rewards	Avg. Win Rates Against Others by	
	Holistic Reward	gpt-3.5-turbo
<i>Task: Long-Form Question Answering</i>		
Factuality	51.93 \pm 1.66	<u>54.74</u> \pm 0.70
Relevance	46.89 \pm 0.72	55.44 \pm 1.12
Completeness	49.51 \pm 0.82	44.97 \pm 1.52
Length	<u>51.67</u> \pm 0.22	44.84 \pm 1.32
<i>Task: Machine Translation*</i>		
Grammar	51.58 \pm 0.64	50.61 \pm 1.62
Confidence	49.20 \pm 0.46	48.66 \pm 1.26
Readability	49.53 \pm 0.42	<u>50.59</u> \pm 1.06
Length	<u>49.69</u> \pm 0.26	50.14 \pm 1.75

Table 7: Averaged win rates against others on both tasks. Each reward is used in ALARM separately. *: Evaluation by gpt-3.5-turbo is on a smaller set of 3K randomly selected examples as in Table 4.

the length reward exploits the flaws. Nonetheless, this is a concern that differs from the inconsistency issues we focus on and falls outside this paper’s scope. See full tables in Appendix A.

Without Combination. To examine whether ALARM helps with more accurate and consistent supervision signals by utilizing both holistic and aspect-specific rewards, we compare the methods using separate rewards individually in the experiments. As shown in Table 1 and 4, ALARM consistently leads to better results along both dimensions.

Without Hierarchical Structure. We contrast our framework with the conventional weighted sum method to highlight the significance of the hierarchical structure. The results from the weighted sum approach reflect a compromise between holistic and aspect-specific rewards, limiting its ability to excel in both. Conversely, our framework, ALARM leverages hierarchical rewards modeling to provide more potent supervision signals, enhancing its performance in both dimensions.

6 Related Work

Hierarchical Reinforcement Learning. Designing a hierarchical rewarding structure that decomposes a long-horizon reinforcement learning task into simpler sub-tasks has shown promising performance in traditional reinforcement learning problems (Florensa et al., 2017; Levy et al., 2019; Kulkarni et al., 2016; Gupta et al., 2020). Motivated by this, ALARM first utilizes hierarchical reinforcement learning to align language models.

Human Preference Alignment. AI alignment with human preference has been one of the key research topics in the NLP community as LLMs show notable performance yet are prone to generating unexpected content (Song et al., 2024; Wang et al., 2023b; Zhou et al., 2023; Yuan et al., 2023). RLHF is a popular algorithm for AI alignment and many related methods are proposed (Dai et al., 2024; Yu et al., 2023; Zhang et al., 2024).

Scalable Oversight. Superhuman models should be capable of handling complex and creative tasks beyond human expertise (Burns et al., 2023), which raises the increasingly important issue of scalable oversight: how to provide reliable supervision signals within limited human capabilities (Amodei et al., 2016; Bowman et al., 2022)? Current methods include improving evaluation quality through human-AI collaboration (Irving et al., 2018; Christiano et al., 2018) and simplifying tasks into sub-tasks for more reliable assessments (Wu et al., 2021; Zhong et al., 2023; Lightman et al., 2023; Liu and Alahi, 2024). Our framework adopts the latter one, identifying simpler aspects for evaluation.

7 Conclusion

We propose ALARM, the first framework hierarchically modeling both holistic and aspect-specific rewards in RLHF. We explore proactive reward selection strategies to enhance compatibility with the holistic reward. The effectiveness of our framework in seeking more accurate and consistent supervision signals and its potential to inspire scalable oversight in AI alignment, is demonstrated through comprehensive experiments, detailed ablation studies, and analyses across two text generation tasks.

Limitations

Our framework requires rewards that are specifically designed for each task, which poses challenges in scaling up the application scenarios. We need to improve the automatic selection of rewards. In our evaluation, we utilize OpenAI’s API, which incurs additional costs and may experience rate limitations and unstable response times.

Ethics Statement

Our study does not involve direct human or animal subjects and presents no discernible ethical concerns. The datasets used, such as QA-Feedback, Europarl, and the toolkits, including Textstat, Lingua, and LanguageTool, are publicly available. We

have taken steps to ensure transparency and reproducibility in our research. We confirm that our research and methodologies are free from harmful practices and potential misuse. We are committed to upholding the highest standards of integrity and ethical responsibility in our work.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62176058) and National Key R&D Program of China (2023YFF1204800). The project’s computational resources are supported by CFFF platform of Fudan University.

References

- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. [Concrete problems in AI safety](#). *CoRR*, abs/1606.06565.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *CoRR*, abs/2204.05862.
- Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. [Longformer: The long-document transformer](#). *CoRR*, abs/2004.05150.
- Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosiute, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McInnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. 2022. [Measuring progress on scalable oversight for large language models](#). *CoRR*, abs/2211.03540.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

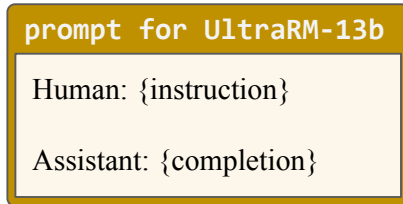
- Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Wojciech Gajewski, Andrea Gesmundo, Neil Houlsby, and Wei Wang. 2018. [Ask the right questions: Active question reformulation with reinforcement learning](#). In *International Conference on Learning Representations*.
- Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. [Weak-to-strong generalization: Eliciting strong capabilities with weak supervision](#). *CoRR*, abs/2312.09390.
- Meng Cao, Lei Shu, Lei Yu, Yun Zhu, Nevan Wichers, Yinxiao Liu, and Lei Meng. 2024. [DRLC: reinforcement learning with dense rewards from LLM critic](#). *CoRR*, abs/2401.07382.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. [Evaluating large language models trained on code](#). *CoRR*, abs/2107.03374.
- Paul F. Christiano, Buck Shlegeris, and Dario Amodei. 2018. [Supervising strong learners by amplifying weak experts](#). *CoRR*, abs/1810.08575.
- Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. [Ultrafeedback: Boosting language models with high-quality feedback](#). *CoRR*, abs/2310.01377.
- Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. [Safe RLHF: Safe reinforcement learning from human feedback](#). In *The Twelfth International Conference on Learning Representations*.
- Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. [AlpacaFarm: A simulation framework for methods that learn from human feedback](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 30039–30069. Curran Associates, Inc.
- Carlos Florensa, Yan Duan, and Pieter Abbeel. 2017. [Stochastic neural networks for hierarchical reinforcement learning](#). In *International Conference on Learning Representations*.
- Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, and Marc Dymetman. 2024. [Compositional preference models for aligning LMs](#). In *The Twelfth International Conference on Learning Representations*.
- Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. 2020. [Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning](#). In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 1025–1037. PMLR.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Geoffrey Irving, Paul F. Christiano, and Dario Amodei. 2018. [AI safety via debate](#). *CoRR*, abs/1805.00899.
- Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. [Personalized soups: Personalized large language model alignment via post-hoc parameter merging](#). *CoRR*, abs/2310.11564.
- Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. [LongEval: Guidelines for human evaluation of faithfulness in long-form summarization](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.
- Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. 2016. [Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation](#). In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- Yuhang Lai, Chengxi Li, Yiming Wang, Tianyi Zhang, Ruiqi Zhong, Luke Zettlemoyer, Wen-Tau Yih,

- Daniel Fried, Sida Wang, and Tao Yu. 2023. [DS-1000: A natural and reliable benchmark for data science code generation](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 18319–18345. PMLR.
- Andrew Levy, Robert Platt, and Kate Saenko. 2019. [Hierarchical reinforcement learning with hindsight](#). In *International Conference on Learning Representations*.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. [Let’s verify step by step](#). *CoRR*, abs/2305.20050.
- Yuejiang Liu and Alexandre Alahi. 2024. [Co-supervised learning: Improving weak-to-strong generalization with hierarchical mixture of experts](#). *CoRR*, abs/2402.15505.
- Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. [Controlling the reading level of machine translation output](#). In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 193–203, Dublin, Ireland. European Association for Machine Translation.
- Ted Moskowitz, Aaditya K Singh, DJ Strouse, Tuomas Sandholm, Ruslan Salakhutdinov, Anca Dragan, and Stephen Marcus McAleer. 2024. [Confronting reward model overoptimization with constrained RLHF](#). In *The Twelfth International Conference on Learning Representations*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Shubham Pateria, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. 2021. [Hierarchical reinforcement learning: A comprehensive survey](#). *ACM Comput. Surv.*, 54(5).
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Seonggi Ryang and Takeshi Abekawa. 2012. [Framework of automatic text summarization using reinforcement learning](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 256–265, Jeju Island, Korea. Association for Computational Linguistics.
- Abdelrhman Saleh, Natasha Jaques, Asma Ghandeharion, Judy Shen, and Rosalind Picard. 2020. [Hierarchical reinforcement learning for open-domain dialog](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8741–8748.
- Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. [Gender bias in machine translation](#). *Transactions of the Association for Computational Linguistics*, 9:845–874.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. [Proximal policy optimization algorithms](#). *CoRR*, abs/1707.06347.
- Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. [Preference ranking optimization for human alignment](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18990–18998.
- Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. [ASQA: Factoid questions meet long-form answers](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. [Learning to summarize with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. [Aligning large multimodal models with factually augmented RLHF](#). *CoRR*, abs/2309.14525.
- Jörg Tiedemann. 2012. [Parallel data, tools and interfaces in OPUS](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten,

- Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and finetuned chat models](#). *CoRR*, abs/2307.09288.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. [A survey on large language model based autonomous agents](#). *Frontiers Comput. Sci.*, 18(6):186345.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. [Large language models are not fair evaluators](#).
- Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2018. Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023b. [Aligning large language models with human: A survey](#).
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. [Finetuned language models are zero-shot learners](#). In *International Conference on Learning Representations*.
- Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul F. Christiano. 2021. [Recursively summarizing books with human feedback](#). *CoRR*, abs/2109.10862.
- Zequ 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](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 59008–59033. Curran Associates, Inc.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Zhaoyang Yang, Kathryn Merrick, Lianwen Jin, and Hussein A. Abbass. 2018. [Hierarchical deep reinforcement learning for continuous action control](#). *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):5174–5184.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2023. [RLHF-V: towards trustworthy mlms via behavior alignment from fine-grained correctional human feedback](#). *CoRR*, abs/2312.00849.
- Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. [RRHF: rank responses to align language models with human feedback without tears](#). *CoRR*, abs/2304.05302.
- Shun Zhang, Zhenfang Chen, Sunli Chen, Yikang Shen, Zhiqing Sun, and Chuang Gan. 2024. [Improving reinforcement learning from human feedback with efficient reward model ensemble](#). *CoRR*, abs/2401.16635.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.
- Ruiqi Zhong, Charlie Victor Snell, Dan Klein, and Jason Eisner. 2023. [Non-programmers can label programs indirectly via active examples: A case study with text-to-SQL](#). In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. [Lima: Less is more for alignment](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

A Experimental Details

We adopt the TRL implementation for the PPO algorithm and conduct all training using LoRA and BF16. The prompt for inferencing UltraRM is shown in Figure 5. We list the resources and materials used in this paper in Table 8.



```
prompt for UltraRM-13b
Human: {instruction}
Assistant: {completion}
```

Figure 5: The inference prompt for UltraRM.

A.1 Training Details on Long-form QA

We limit the max input length to 1024 and the max output length to 200 following the original setting. We set the weight of the holistic reward to 5 through trials and keep other weights to 1. The factuality reward model predicts 0.5 and -0.5 for the correct and wrong sentences respectively. The batch size is set to 8 and the mini-batch size to 4, with gradient accumulation steps set to 2. We set the learning rate to $2e - 4$. Other PPO hyper-parameters should follow the default values of TRL implementation. The training takes around 2 hours on 8x48G NVIDIA A6000 GPU.

A.2 Training Details on MT

We set the max length to 128 for both input and output. We set the weight of the holistic reward to 3 through trials and keep other weights to 1. The grammar reward is designed to predict -1 for those incorrect tokens and 0 for the correct ones before the reward shaping. The batch size is set to 32 and the mini batch size is set to 16, with gradient accumulation steps set to 2. We set the learning rate to $5e - 4$. Other PPO hyper-parameters also follow the default values. The training takes around 1 hour on 8x48G NVIDIA A6000 GPU.

A.3 Additional Ablation Study Results on Reward Selection

We put the full results of win rates in the ablation study for reward selection in Table 9 and Table 10. For those extensive experiments, We use the same settings as the main ones. The relevance model predicts 0.3 and -0.3 for the correct and wrong sentences respectively. We apply z-normalization

to the readability and completeness rewards. We divide the token count of a generation by the average token count on the training set and define the value as its length reward. The language confidence remains the same for rewarding since it has a suitable value range between 0 and 1.

B Evaluation Details

B.1 GPT as an Evaluator

We use the gpt-3.5-turbo-1106 version of OpenAI API for pairwise evaluation. Figure 6 shows our prompt to call gpt-3.5-turbo in pairwise comparisons for both tasks, which is originally from AlpacaEval (Dubois et al., 2023).

B.2 Statistical Testing

We aggregate the data from all runs and then compute the p-values for the mean rewards using the paired t-test function from SciPy. Similarly, we determine the p-values for win rates by employing the binomial test function from SciPy. We interpret a p-value greater than 0.05 as a statistical tie.

Resources and Materials	Access Link
UltraRM-13b	https://huggingface.co/openbmb/UltraRM-13b
Long-form QA	https://github.com/allenai/FineGrainedRLHF
Europarl	https://huggingface.co/datasets/Helsinki-NLP/europarl
LanguageTool	https://github.com/language-tool-org/language-tool
Lingua	https://github.com/pemistahl/lingua-py
textstat	https://github.com/textstat/textstat
TRL	https://github.com/huggingface/trl

Table 8: The resources and materials we utilize in this paper.

Rewards	Factuality	Relevance	Completeness	Length	Avg.
<i>% Win Rates by the Holistic Reward</i>					
Factuality	-	53.57 ± 2.06	51.74 ± 1.64	50.49 ± 1.58	51.93 ± 1.66
Relevance	46.43 ± 2.06	-	47.88 ± 0.49	46.36 ± 0.30	46.89 ± 0.72
Completeness	48.26 ± 1.64	52.12 ± 0.49	-	48.16 ± 1.33	49.51 ± 0.82
Length	49.51 ± 1.58	53.64 ± 0.30	51.83 ± 1.33	-	<u>51.67</u> ± 0.22
<i>% Win Rates Evaluated by gpt-3.5-turbo</i>					
Factuality	-	48.95 ± 1.01	56.57 ± 1.89	58.70 ± 0.75	<u>54.74</u> ± 0.70
Relevance	51.05 ± 1.01	-	58.33 ± 0.64	56.96 ± 1.79	55.44 ± 1.12
Completeness	43.43 ± 1.89	41.67 ± 0.64	-	49.81 ± 2.44	44.97 ± 1.52
Length	41.30 ± 0.75	43.04 ± 1.79	50.19 ± 2.44	-	44.84 ± 1.32

Table 9: Additional ablation study results of win rates in long-form QA.

Rewards	Grammar	Confidence	Readability	Length	Avg.
<i>% Win Rates by the Holistic Reward</i>					
Grammar	-	51.79 ± 0.85	51.60 ± 0.78	51.36 ± 0.30	51.58 ± 0.64
Confidence	48.21 ± 0.85	-	49.79 ± 0.15	49.60 ± 0.44	49.20 ± 0.46
Readability	48.40 ± 0.78	50.21 ± 0.15	-	49.98 ± 0.57	49.53 ± 0.42
Length	48.64 ± 0.30	50.40 ± 0.44	50.02 ± 0.57	-	<u>49.69</u> ± 0.26
<i>% Win Rates* Evaluated by gpt-3.5-turbo</i>					
Grammar	-	51.31 ± 2.11	50.19 ± 0.44	50.34 ± 2.50	50.61 ± 1.62
Confidence	48.69 ± 2.11	-	48.41 ± 1.71	48.87 ± 1.19	48.66 ± 1.26
Readability	49.81 ± 0.44	51.59 ± 1.71	-	50.38 ± 1.95	<u>50.59</u> ± 1.06
Length	49.66 ± 2.50	51.13 ± 1.19	49.62 ± 1.95	-	50.14 ± 1.75

Table 10: Additional ablation study results of win rates in MT. *: Evaluation is on a smaller set of 3K randomly selected examples as in Table 4.

prompt for gpt-3.5-turbo-1106

```
<|im_start|>system
You are a helpful instruction-following assistant that prints the best model by selecting the best
outputs for a given instruction.
<|im_end|>
<|im_start|>user
Select the output (a) or (b) that best matches the given instruction. Choose your preferred
output, which can be subjective. Your answer should ONLY contain: Output (a) or Output (b).
Here's an example:

# Example:
## Instruction:
Give a description of the following job: "ophthalmologist"

## Output (a):
An ophthalmologist is a medical doctor who pokes and prods at your eyes while asking you to
read letters from a chart.

## Output (b):
An ophthalmologist is a medical doctor who specializes in the diagnosis and treatment of eye
diseases and conditions.

## Which is best, Output (a) or Output (b)?
Output (b)

Here the answer is Output (b) because it provides a comprehensive and accurate description of
the job of an ophthalmologist. In contrast, output (a) is more of a joke.

# Task:
Now is the real task, do not explain your answer, just say Output (a) or Output (b).

## Instruction:
{instruction}

## Output (a):
{output_1}

## Output (b):
{output_2}

## Which is best, Output (a) or Output (b)?
<|im_end|>
```

Figure 6: The evaluation prompt for gpt-3.5-turbo-1106.