Curiosity-Driven Reinforcement Learning from Human Feedback

Haoran Sun* **Yekun Chai*♠** **Shuohuan Wang**
Yu Sun **Hua Wu** **Haifeng Wang**
Baidu Inc.
sunhaoran0402@gmail.com
{chaiyekun,wangshuohuan}@baidu.com

Abstract

Reinforcement learning from human feedback (RLHF) has proven effective in aligning large language models (LLMs) with human preferences, but often at the cost of reduced output diversity. This trade-off between diversity and alignment quality remains a significant challenge. Drawing inspiration from curiosity-driven exploration in reinforcement learning (Pathak et al., 2017), we introduce curiosity-driven RLHF (CD-RLHF), a framework that incorporates intrinsic rewards for novel states, alongside traditional sparse extrinsic rewards, to optimize both output diversity and alignment quality. We demonstrate the effectiveness of CD-RLHF through extensive experiments on a range of tasks, including text summarization and instruction following. Our approach achieves significant gains in diversity on multiple diversity-oriented metrics while maintaining alignment with human preferences comparable to standard RLHF. We make our code publicly available at https://github.com/ernie-research/CD-RLHF.

1 Introduction

Reinforcement learning from human feedback (RLHF) (Ziegler et al., 2019; Christiano et al., 2017) is now a critical component in fine-tuning large language models (LLMs) to specific tasks like code generation (Roziere et al., 2023; Chai et al., 2023; Lozhkov et al., 2024), mathematical reasoning (Uesato et al., 2022), and dialogue assistance (Ouyang et al., 2022; Askell et al., 2021).

Despite the success of RLHF in producing high-performing LLMs, it often reduces the output diversity of LLMs (Dong et al., 2023; Kirk et al., 2024). Kirk et al. (2024); Wu et al. (2024) find out that models trained with RLHF exhibit a *trade-off* between alignment quality and output diversity: RLHF models have high alignment quality but with low output diversity, and vice versa. LLMs with decreased output diversity could potentially hinder the LLMs' effectiveness on creative and open-ended tasks, like story generation (Castricato et al., 2022; Yang et al., 2024), data synthesis (Xu et al., 2023; Liang et al., 2024; Dubey et al., 2024), and red-teaming (Perez et al., 2022; Hong et al., 2024).

Some attempts have been made to improve the diversity or achieve a balance in this trade-off. Hong et al. (2024) introduce SelfBLEU and Sentence-BERT metrics into the RL training as sparse rewards, mainly towards a coverage of test cases in red-teaming. Moreover, Wang et al. (2024) gives a solution from the view of Kullback-Leibler (KL) penalty, implementing DPO or PPO with forward KL gains output diversity, but sacrifices the alignment quality. Therefore, is it possible to balance this trade-off in RLHF stage, thus improving output diversity *without* sacrificing alignment quality?

Curiosity-driven exploration methods have been well-studied in reinforcement learning (RL) literatures (Schmidhuber, 1991b; Bellemare et al., 2016; Ostrovski et al., 2017; Pathak et al., 2017; Burda et al., 2019b). In these RL settings, agents (*i.e.,* policy models) are encouraged to explore the "novel" states, which are less visited in the learning progress. The curiosity signal introduced in this framework serves as a metric to estimate the "novelty" of a state and is implemented as an intrinsic reward. Under this setting, agents explore the "novel" states with curiosity, learning skills that may used to solve the latter problems.

Inspired by these studies, we propose curiosity-driven reinforcement learning from human feedback (CD-RLHF), a novel framework that encourages agents to explore more often with curiosity in RLHF stage. As curiosity-driven RL, curiosities are implemented as intrinsic rewards and assigned along the trajectory. Combining intrinsic rewards with the extrinsic rewards generated by the reward

*Equal contribution and shared co-first authorship. Work done during HS's internship at Baidu.
♠Correspondence to: YC.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23517–23534
July 27 - August 1, 2025

model helps guide the agent policy in optimizing both output diversity and alignment quality. In CD-RLHF, we use forward dynamics to compute the prediction errors in the state representations at each time and use these errors to estimate curiosity, following Pathak et al. (2017); Burda et al. (2019b). This estimation shows a good property: with the training progress, a frequently visited "novel" state becomes boring and is less valuable for further exploration. Using curiosity encourages the agent to make various token choices in the same state potentially leading to the improvement of output diversity. Besides, the guidance of extrinsic reward ensures that the agent is still optimized towards alignment quality.

Contribution To conclude, our main contributions are summarized as follows:

- We propose curiosity-driven RLHF, a framework that maintains alignment performance comparable to RLHF while achieving higher output diversity. We demonstrate its effectiveness through extensive experiments across various datasets and tasks, including text summarization and instruction following.
- We demonstrate that CD-RLHF improves output diversity evaluated by diversity-oriented metrics in both lexical and semantic dimensions, while preserving alignment quality comparable to vanilla RLHF.
- In the out-of-distribution benchmark MT-Bench, CD-RLHF demonstrates strong generalization capability, achieving higher MT-Bench scores compared to RLHF evaluated by GPT-4, and consistently showing more diversity outputs.

2 Related Work

LLM Alignment RLHF significantly enhances the alignment of LLMs, improving instruction following, safety, and reliability in real-world applications (Ziegler et al., 2019; Bai et al., 2022; Li et al., 2024). Recent efforts in efficient alignment algorithms focus on general performance improvement (Li et al., 2023; Ahmadian et al., 2024; Chai et al., 2024), directly aligning with preference data (Rafailov et al., 2024; Xu et al., 2024; Gheshlaghi Azar et al., 2023), and tacking credit assignment with dense rewards (Wu et al., 2023; Chan et al., 2024; Yoon et al., 2024). However, alignment can degrade output diversity, as pointed out by Kirk et al. (2024) who found that RLHF-trained LLMs show less diversity than supervised fine-tuned models. Nevertheless, Wu et al. (2024) defines this as generative monoculture, which means LLMs refuse to share diverse options. There are some attempts to try to mitigate this situation with different approaches. Wang et al. (2024) proposes using f-divergence with DPO/PPO to improve diversity at the cost of alignment, different divergences show a different level of trade-off between diversity and alignment. Hong et al. (2024) integrate SelfBLEU and Sentence-BERT into RL fine-tuning, boosting output diversity for the improvement of coverage of test cases in red-teaming tasks. Bradley et al. (2023) proposes a method that is applied at the inference stage, generating a new solution based a previous solution. These solutions are evaluated with LLMs considering both quality and diversity, and the better one is reserved.

Curiosity-Driven RL Curiosity-driven approaches have been widely studied in RL. Count-based approaches are successful in tabular settings but fail at generalising to additional tasks with infinite states (Bellemare et al., 2016; Ostrovski et al., 2017). Using prediction error as curiosity approaches highlight their success in multiple RL scenarios (Schmidhuber, 1991b,a; Stadie et al., 2015). Specifically, methods proposed by (Pathak et al., 2017; Burda et al., 2019a,b) treat curiosities as intrinsic rewards, and use intrinsic rewards to encourage agents to explore at novel states, intending to enhance agent capabilities that could be used in future environments. Pure curiosity-driven learning has also been studied in large-scale settings (Burda et al., 2019a).

3 Method

In this section, we introduce our CD-RLHF method. The key idea is to encourage the policy to explore states with high curiosity potential, which may yield a broader range of diverse tokens and thus improve output diversity.

3.1 Preliminaries: Curiosity-Driven RL

Curiosity-driven RL (Pathak et al., 2017; Burda et al., 2019b), also known as exploration bonus, encourages an agent to explore novel states with curiosities that serve as the intrinsic rewards $r^{(i)}$. By doing so, the sparse extrinsic reward $r^{(e)}$ which guides the agent from a holistic perspective is replaced with a new reward $r_t = r_t^{(i)} + r_t^{(e)}$. The $r_t^{(e)}$ is mostly near zero along the trajectory, except for the end of the episode.

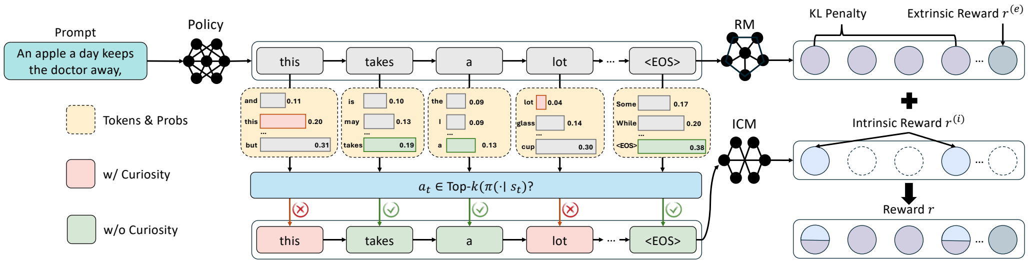


Figure 1: Illustration of the CD-RLHF framework. The policy model generates a completion based on the given instruction, which samples tokens from vocabulary at each time. The introduced intrinsic curiosity module (ICM) estimates the curiosity as a metric for "novelty" of the context, producing the intrinsic rewards. Another mechanism is introduced to determine which context is worth exploring, based on whether the selected token's probability ranks among the top-k highest probabilities ($k = 1$ in the illustration).

The motivation behind curiosity-driven RL is that intrinsic rewards should be higher in novel states than in normal states to encourage the agent to visit. A common approach in the RL literature is to estimate intrinsic rewards (*i.e.*, curiosity) using prediction errors. The forward dynamics (Schmidhuber, 1991b; Pathak et al., 2017; Burda et al., 2019b) or inverse dynamics (Haber et al., 2018) are utilized to predict the representation of the next state, then compute the prediction error between the predicted and ground truth representation of the next state. By maximizing the prediction errors in RL procedure, agents tend to get attracted to the stochastic transitions, which correspond to novel states. During the training progress, the prediction errors tend to decrease as the agent becomes familiar with more states. Since the prediction happens in the latent space, the choice of features to estimate the prediction error is a vital factor. As indicated by Burda et al. (2019a), the latent variables generalise better than the raw observations (*i.e.*, pixel), making them more effective for estimating the intrinsic rewards.

3.2 Curiosity-Driven Alignment

Our insight is that the curiosity in RLHF reflects the agent's unfamiliarity with the state. By frequently exploring such state with high curiosity, the agent is encouraged to select diverse actions under the same state across trajectories, thereby enhancing output diversity. In CD-RLHF, we integrate a intrinsic curiosity module (ICM) into RLHF framework to estimate this curiosity. The estimation happens only at the states with enough "novelty", which is determined by the probability of the selected token, as shown in Figure 1.

3.2.1 Curiosity-Driven Reward

The extrinsic reward $r_t^{(e)}$ at time step t is produced by an external reward model, and obtained at the end of the episode. In RLHF, the Kullback-Leibler (KL) penalty works as a token-level reward, which constrains the policy's divergence from the supervised fine-tuned (reference) model:

$$r^{(e)} = R - \beta D_{\text{KL}}(\pi_{\text{policy}}(\cdot) \parallel \pi_{\text{ref}}(\cdot))$$

where R is the reward produced by the reward model, π_{policy} is the policy model and π_{ref} is the reference model.

Combining intrinsic reward and extrinsic reward, we can obtain the reward that is used to optimize the policy in CD-RLHF:

$$r_t = r_t^{(e)} + \eta \cdot r_t^{(i)}$$

where the discounting factor η controls the scale of intrinsic rewards.

This reward is designed to steer policy optimization toward achieving both alignment quality, driven by extrinsic rewards, and output diversity, driven by intrinsic rewards.

3.2.2 Prediction Error as Curiosity Reward

In the RLHF framework, the state consists of a sequence of action tokens up to current time step, and the action selected at each step represents the decision made. At time step t, the action is denoted as a_t, and the state is represented as $s_t = \{s_0, a_{<t}\}$, where s_0 is the initial prompt and $a_{<t}$ refers to the sequence of actions taken before time t. The transition function can be interpreted as updating the state with the generated token at each time step, where the state at time $t+1$ is represented as $s_{t+1} = \{s_t, a_t\}$, reflecting the concatenation of the previous state s_t and the newly generated token a_t.

Intrinsic Curiosity Module (ICM) To estimate intrinsic rewards, we introduce an Intrinsic Curiosity Module (ICM), inspired by Pathak et al. (2017); Burda et al. (2019b). The ICM predicts the next state representation based on the current state and action, and the prediction error serves as the curiosity-driven intrinsic reward. This prediction error is computed in the latent space of hidden embeddings from large language models (LLMs), enabling a more expressive and informative representation of states and actions.

ICM consists of a feature encoder ϕ and a forward model f, both implemented as two-layer MLPs. The feature encoder ϕ encodes the state representation to bridge the scalar gap between raw state and action representations. The forward model f takes as input the concatenation of the encoded state $\phi(s_t)$ and the action representation $\psi(a_t)$, predicting the next state representation $\phi(s_{t+1})$. This prediction process is formalized as:

$$\hat{\phi}(s_{t+1}) = f(\phi(s_t), \psi(a_t)) \quad (1)$$

The ICM loss $\mathcal{L}_{\text{ICM}}$ is computed as the squared error between the predicted and actual next-state representations:

$$\mathcal{L}_{\text{ICM}} = \frac{1}{2}\|\hat{\phi}(s_{t+1}) - \phi(s_{t+1})\|_2^2 \quad (2)$$

where $\hat{\phi}(s_{t+1})$ is the predicted state representation, $\phi(s_{t+1})$ is the true latent representation.

Intrinsic Reward Computation The ICM is trained independently of the policy model, optimizing $\mathcal{L}_{\text{ICM}}$ as a self-supervised learning task. The resulting prediction error is used as the intrinsic reward, encouraging the exploration of less predictable states. Intuitively, deterministic states with limited action choices offer little exploratory value. To improve the efficiency, we apply a top-k strategy to determine whether a state warrants further exploration. The intrinsic reward $r_t^{(i)}$ at time step $t+1$ is defined as:

$$r_t^{(i)} = \begin{cases} 0 \text{ if } a_t \in \text{Top-}k(\pi(\cdot \mid s_t)) \\ \frac{1}{2}\|\hat{\phi}(s_{t+1}) - \phi(s_{t+1})\|_2 \text{ otherwise} \end{cases}$$

Here, Top-$k(\pi(\cdot \mid s_t))$ denotes the top-k vocabulary subset that maximizes $\sum_{a_t \in \text{Top-}k(\cdot)} \pi(a_t \mid s_t)$. When the selected action a_t is within the top-k most probable candidates, the intrinsic reward is set to zero to discourage redundant exploration.

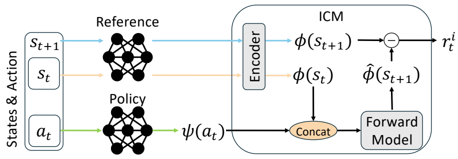


Figure 2: The computation of ICM loss $\mathcal{L}_{\text{ICM}}$ and intrinsic reward $r_t^{(i)}$ at time t, the state representation s_t and s_{t+1} are derived from the reference model, while the action representation a_t is taken from the policy model.

The intrinsic reward computation aligns with $\mathcal{L}_{\text{ICM}}$, as the state prediction accuracy improves during ICM training, reducing the intrinsic reward over time. This follows the principle of curiosity-driven RL—novel states become less rewarding with frequent visits. Figure 2 illustrates the intrinsic reward estimation process.

Reward Whitening The scalar gap between intrinsic and extrinsic rewards could introduce unstable optimization in RL training, where intrinsic rewards may vary across samples. To mitigate this, we apply a *reward whitening* to normalize intrinsic rewards at each time step along the trajectory:

$$r^{(i)} = (r^{(i)} - \mu)/\sigma^2$$

where μ and σ represent the mean and standard deviation of the intrinsic rewards, respectively.

Feature Space Alignment During the intrinsic reward estimation in RLHF, there is a gap between predictions and the states/actions. Predictions are operated in continuous space, while states and actions reside in discrete space. To tackle this gap, we use the last layers' hidden states s_t derived from the reference model, and use the embeddings of the actor model to extract the representation of a_t. This alignment results in a coherent feature space for computing prediction errors.

4 Experiments

4.1 Experimental Settings

Datasets We evaluate our method on the TL;DR (Stiennon et al., 2020) dataset for text summarization, which includes 93k human-annotated preference pairs and 86k pairs for evaluation. Additionally, we evaluate on the UltraFeedback (Cui et al., 2023) dataset for instruction following, which contains 61.1k pairs for training and 2k pairs for evaluation.[1]

[1]We use the binarized version from https://huggingface.co/datasets/HuggingFaceH4/UltraFeedback_binarized, following the preference modeling splits in all experiments.

Models and training details We use pre-trained Gemma-2B, Gemma-7B (Team et al., 2024), Llama-3.2-1B, and Llama-3.2-3B (Dubey et al., 2024) as base models. CD-RLHF is implemented using the DeepSpeed-Chat (Yao et al., 2023) framework. The datasets are split into SFT/RM/PPO partitions with a 20%/40%/40% ratio, respectively. Hyperparameters used for SFT, RM, and PPO stages are detailed in Appendix A. In CD-RLHF, we use top-1 to select valid intrinsic rewards.

Compared baselines In addition to the vanilla RLHF baseline, we evaluate other approaches aimed at enhancing output diversity. Hong et al. (2024) propose optimizing SelfBLEU and Sentence-BERT scores across samples during RLHF training as sentence-level additional rewards, and optimizing the entropy as token-level additional rewards. We denote this method as Sent-Rewards and re-implement it within DeepSpeed-Chat, using consistent hyperparameters in PPO training for fair comparison. Additional details of implementing Sent-Rewards can be found in Appendix A.

Evaluation We evaluate performance using both quality and diversity metrics.

- **RM scores**: RM scores evaluate the alignment quality using the trained reward model, which is identical to the one used in PPO training.
- N**-gram distinct** (Li et al., 2016): It measures the number of distinct N-grams in total N-grams (averaging 1- to 5-grams). This metric is referred as Diversity.
- **EAD** (Liu et al., 2022): EAD is a refined version of N-gram Distinct, EAD adjust for length bias by introducing vocabulary size.
- **SelfBLEU** (Zhu et al., 2018): It calculates BLEU scores within outputs, by taking one sentence as hypothesis and the rest as reference to assess lexical diversity (averaging 1- to 4-grams).
- **Sentence-BERT** (Reimers and Gurevych, 2019): It computes semantic diversity by averaging cosine similarity between output embeddings. We refer to this metric as SentBERT.
- **GPT-4 win-rate** (Tevet and Berant, 2021): We utilize GPT-4 to evaluate the diversity by choosing the more diversity set from two sets of completions, which are produced by different models.
- **Human win-rate**: Similar to GPT-4 win-rate, we assess the win-rates between two methods with the assistance of human annotators.

For the automated metrics, we randomly sample 2k instances from each dataset's validation/test set, generating 10 completions for an instruction with a temperature of 1.0. For diversity related metrics, we compute the per-input diversity which estimate the diversity within the 10 completions for an instruction, and averaged across instances. We sampled 50 instances from the 2k subset to assess the win-rate estimated by GPT-4. To decrease the evaluation difficulty of human evaluation, we sampled 25 instances from GPT-4 evaluation subset, with each instruction has 5 completions.

4.2 Main Results

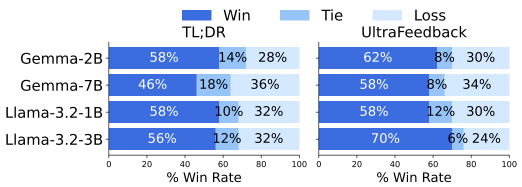
The evaluation results are shown in Table 1. Overall, CD-RLHF consistently outperforms the PPO baseline and Sent-Rewards in terms of output diversity, while achieving a favorable balance between output diversity and alignment quality.

Results on TL;DR For the TL;DR summarization task, both CD-RLHF and Sent-Rewards improve output diversity over the PPO baseline. CD-RLHF achieving the highest diversity scores among all RLHF models, obtains 16.66% improvements on Gemma-2B and 6.22% on Gemma-7B compared to RLHF baseline. Additionally, CD-RLHF maintains comparable alignment quality (RM scores) to the PPO baseline. For Llama-3.2-1B and Llama-3.2-3B, CD-RLHF achieves remarkable performance gain on both output diversity, with 18.30% and 14.41% improvements on output diversity in average, respectively. Besides, CD-RLHF on Llama-3.2-3B brings improvements on alignment quality with +0.16 in RM scores.

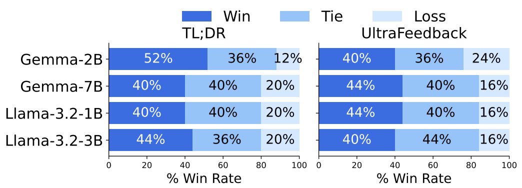
Results on UltraFeedback For the UltraFeedback instruction-following task, CD-RLHF demonstrates substantial improvements in output diversity for all models. While the Sent-Rewards method helps mitigate output diversity degradation, it remains less effective than CD-RLHF in balancing this trade-off. In this setting, CD-RLHF effectively manages the trade-off, showing only a minimal drop in quality on the Gemma-7B (-0.01 in RM score) and Llama-3.2-1B (-0.03 in RM score). Compared to RLHF baseline, CD-RLHF achieves an average of 9.84% on Gemma-2B, 8.28% on Gemma-7B, 7.35% on Llama-3.2-1B, and 14.29% on Llama-3.2-3B. Additionally, CD-RLHF outperforms the SFT model in lexical diversity, achieving higher scores on Diversity, EAD, and SelfBLEU

Stage	Model	Diversity ↑	EAD ↑	SelfBLEU ↓	SentBERT ↓	RM Score ↑
TL;DR						
SFT	Gemma-2B	0.4041	0.8444	0.1336	0.5380	-0.52
RLHF		0.2132	0.7347	0.3367	0.7024	0.90
Sent-Rewards (Hong et al., 2024)		0.2355	0.7512	0.3053	0.6961	**0.95**
CD-RLHF		**0.2839**	**0.7793**	**0.2590**	**0.6720**	**0.95**
(Δ over RLHF)		**(+33.16%)**	**(+6.07%)**	**(+23.08%)**	**(+4.33%)**	
SFT	Gemma-7B	0.3575	0.8256	0.1706	0.5442	0.62
RLHF		0.1180	0.6602	0.4352	0.7601	**2.02**
Sent-Rewards (Hong et al., 2024)		**0.1447**	0.6735	0.4197	0.7749	1.87
CD-RLHF		0.1360	**0.6816**	**0.4144**	**0.7480**	**2.02**
(Δ over RLHF)		**(+15.25%)**	**(+3.24%)**	**(+4.78%)**	**(+1.59%)**	
SFT	Llama-3.2-1B	0.3672	0.8296	0.1651	0.5498	-0.31
RLHF		0.1724	0.6869	0.3997	0.6971	1.14
Sent-Rewards (Hong et al., 2024)		0.1869	0.7078	0.3519	0.7338	0.80
CD-RLHF		**0.2418**	**0.7482**	**0.3108**	**0.6847**	**1.17**
(Δ over RLHF)		**(+40.26%)**	**(+8.92%)**	**(+22.24%)**	**(+1.78%)**	
SFT	Llama-3.2-3B	0.3539	0.8233	0.1753	0.5649	2.32
RLHF		0.2281	0.7441	0.3163	0.6658	3.33
Sent-Rewards (Hong et al., 2024)		0.2355	0.7530	0.3356	0.7182	**3.54**
CD-RLHF		**0.2920**	**0.7879**	**0.2463**	**0.6551**	3.49
(Δ over RLHF)		**(+28.01%)**	**(+5.89%)**	**(+22.13%)**	**(+1.61%)**	
UltraFeedback						
SFT	Gemma-2B	0.1855	0.7346	0.2899	0.6766	-1.45
RLHF		0.1686	0.6503	0.3104	0.7672	-1.01
Sent-Rewards (Hong et al., 2024)		0.1603	0.6801	0.3483	0.7463	**-0.80**
CD-RLHF		**0.1899**	**0.7417**	**0.2858**	**0.7308**	-0.90
(Δ over RLHF)		**(+12.63%)**	**(+14.06%)**	**(+7.93%)**	**(+4.74%)**	
SFT	Gemma-7B	0.2148	0.7594	0.2292	0.6378	0.29
RLHF		0.2345	0.7360	0.2717	0.7298	**0.63**
Sent-Rewards (Hong et al., 2024)		0.2557	**0.7744**	0.2512	0.7231	0.60
CD-RLHF		**0.2654**	0.7639	**0.2442**	**0.6858**	0.62
(Δ over RLHF)		**(+13.18%)**	**(+3.79%)**	**(+10.12%)**	**(+6.03%)**	
SFT	Llama-3.2-1B	0.2060	0.7545	0.2476	0.6517	-0.03
RLHF		0.1683	0.6499	0.3564	0.7813	**1.00**
Sent-Rewards (Hong et al., 2024)		0.1652	0.6170	0.3345	0.7924	0.97
CD-RLHF		**0.1834**	**0.6891**	**0.3149**	**0.7598**	0.97
(Δ over RLHF)		**(+8.97%)**	**(+6.03%)**	**(+11.64%)**	**(+2.75%)**	
SFT	Llama-3.2-3B	0.1974	0.7484	0.2669	0.6885	0.59
RLHF		0.1805	0.7031	0.3188	0.7676	1.35
Sent-Rewards (Hong et al., 2024)		0.1765	0.7261	0.2944	0.7471	1.22
CD-RLHF		**0.2223**	**0.7673**	**0.2531**	**0.7349**	**1.43**
(Δ over RLHF)		**(+23.16%)**	**(+9.13%)**	**(+20.61%)**	**(+4.26%)**	

Table 1: Results on TL;DR and UltraFeedback datasets, comparing CD-RLHF with SFT, RLHF, and Sent-Rewards (Hong et al., 2024) from both output diversity and alignment quality aspects.



(a) GPT-4 Win-Rates



(b) Human Win-Rates

Figure 3: The win rates of CD-RLHF against vanilla PPO when evaluating output diversity on TL;DR and UltraFeedback datasets.

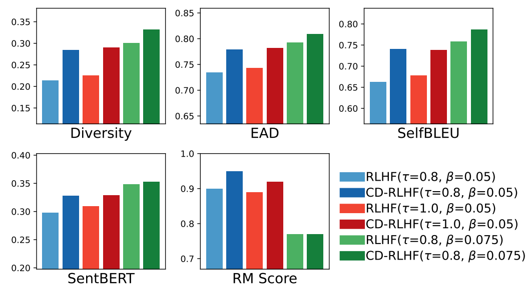


Figure 4: Output diversity and alignment performance of RLHF and CD-RLHF under different hyperparameter settings: KL coefficient β and sampling temperature τ. For a better illustration, we use 1.0 - SelfBLEU and 1.0 - SentBERT as metric to maintain a similar trend with other metrics.

metrics for the Gemma-2B and Llama-3.2-3B, and on Diversity and EAD metrics for the Gemma-7B.

GPT-4 and human evaluation For further evaluation using GPT-4, as shown in Figure 3(a), CD-RLHF achieves a win rate of 58%, 46%, 58% and 56% over the PPO baseline on the TL;DR dataset for the Gemma-2B, Gemma-7B, Llama-3.2-1B and Llama-3.2-3B, respectively. On the UltraFeedback dataset, CD-RLHF consistently yields an average 62% win rates across the evaluated models. In Figure 3(b), we demonstrate the human evaluations on the TL;DR and UltraFeedback datasets, where

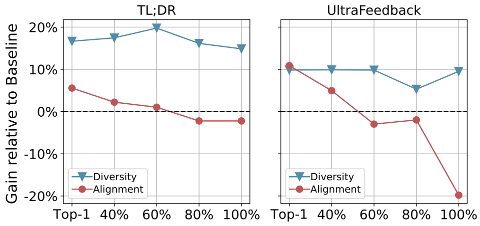


Figure 5: Diversity and alignment quality changes with increasing the frequency of intrinsic rewards. The y-axis represents the improvement ratio against RLHF baseline. The diversity is the average of the related metrics: Diversity, EAD, SelfBLEU, and SentBERT, and alignment is the RM score.

CD-RLHF consistently outperforms RLHF on all evaluated models. These comparison evaluation results demonstrate the remarkable performance CD-RLHF achieved in improving output diversity.

4.3 Ablation Studies

In this section, we conduct ablation studies on Gemma-2B for better analysis with various hyper-parameter settings.

As previous works (Kirk et al., 2024; Wang et al., 2024) have shown, the KL divergence in RLHF heavily influences the balance between output diversity and alignment performance. In Figure 4 (blue and red bars), we investigate CD-RLHF's performance under various KL settings: KL coefficient $\beta = 0.05$ and 0.075. Within each method, we observe that as β increases, output diversity improves, albeit at the cost of alignment performance. However, under the same KL setting, CD-RLHF consistently outperforms RLHF in output diversity, without compromising alignment quality.

Another factor that could affect the performance of CD-RLHF is the choice of the sampling hyper-parameters. In this section, we examine the effect of sampling with different settings: temperature $\tau = 0.8$ and 1.0. Figure 4 (blue and green bars) shows the performance of RLHF and CD-RLHF under these sampling temperatures during training. While increasing the sampling temperature during experience generation moderately improves output diversity, its impact is less pronounced than that of increasing the KL. Under both temperature settings, CD-RLHF still show improvements on output diversity, and also the alignment quality.

4.4 Analysis

Frequency of intrinsic rewards In the computation described in §3.2.2, the intrinsic rewards are valid only when the probability of the selected token lies out of top-k candidates. This raises the question of whether this subset is adequate for curiosity exploration. To investigate, we expand the frequency of intrinsic rewards from top-1 ($\approx 20\%$) to 40%, 60%, 80% and 100% of the length of trajectory, selected at random.

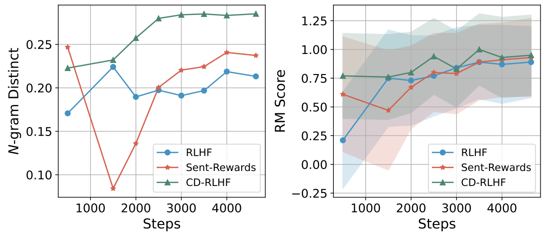


Figure 6: N-gram distinct (diversity) and RM score (alignment quality) curves during training for RLHF, Sent-Rewards, and CD-RLHF methods.

Figure 5 shows the impact on output diversity as the frequency of intrinsic rewards increases. For the TL;DR summarization task, increasing the ratio to 60% leads to a modest improvement in output diversity, with a 3% average increase. However, when expanding to cover 100%, output diversity remains similar to the 20% case, but alignment quality declines from 0.95 to 0.88. A similar trend is observed on the UltraFeedback dataset: increasing the frequency of intrinsic rewards provides comparable output diversity to the top-1 setting, yet significantly reduces alignment quality. These findings suggest that focusing on tokens beyond the top-1 probability is sufficient to identify the contexts in which curiosity-driven exploration is beneficial.

Training curves of diversity and alignment quality We compare diversity and alignment quality changes during training for RLHF, CD-RLHF, and Sent-Rewards on the TL;DR task using Gemma-2B model. We use N-gram distinct to measure diversity and the RM score to assess alignment quality, with results shown in Figure 6.

For diversity, we observe that RLHF maintains a relatively low diversity level throughout training, after an initial decrease from the SFT model. Both CD-RLHF and Sent-Rewards positively impact diversity, though CD-RLHF consistently improves diversity throughout training, while Sent-Rewards experiences degradation until around 1500 steps.

Regarding alignment quality, all methods show positive effects during training. Notably, CD-RLHF converges faster than the other two, reaching comparable performance at step 2500 rather than at step 4640 (the end of training). Additionally, CD-RLHF displays a more stable alignment quality curve than Sent-Rewards.

		MT-Bench Scores ↑			Diversity Metrics			
Stage	**Model**	**Turn 1**	**Turn 2**	**Overall**	**Diversity ↑**	**EAD ↑**	**SelfBLEU ↓**	**SentBERT ↓**
RLHF	Gemma-2B	6.26	**4.45**	5.35	0.1076	0.6160	0.4383	0.7592
CD-RLHF		**6.91**	**4.45**	**5.68**	**0.1123**	**0.6292**	**0.3961**	**0.7315**
RLHF	Gemma-7B	6.36	5.15	5.75	0.1173	0.5850	0.4851	0.7854
CD-RLHF		**6.46**	**5.46**	**5.96**	**0.1297**	**0.6051**	**0.4623**	**0.7617**
RLHF	Llama-3.2-1B	4.33	3.10	3.71	0.0818	0.5356	0.4699	0.7475
CD-RLHF		**4.78**	**3.57**	**4.18**	**0.0895**	**0.5629**	**0.3919**	**0.7282**
RLHF	Llama-3.2-3B	6.47	**5.47**	5.98	0.0939	0.5878	0.4489	0.7797
CD-RLHF		**6.71**	5.45	**6.08**	**0.1133**	**0.6213**	**0.3962**	**0.7529**

Table 2: Results on MT-Bench with MT-Bench scores estimate quality of response, and diversity metrics estimate output diversity. All of the models are trained on the UltraFeedback dataset.

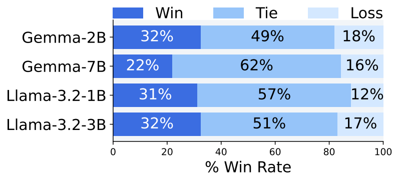


Figure 7: Win rates of CD-RLHF against vanilla RLHF on MT-Bench, estimated by GPT-4.

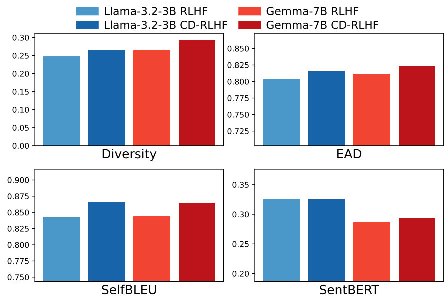


Figure 8: Output diversity of RLHF and CD-RLHF on story writing task. For a better illustration, we use 1.0 - SelfBLEU and 1.0 - SentBERT as metric to maintain a similar trend with other metrics. The higher the better.

4.5 Out-of-Distribution Analysis

In this section, we explore whether the increased output diversity achieved with CD-RLHF impacts generalization capability or potentially enhances it. We use the widely adopted MT-Bench (Zheng et al., 2024) as an out-of-distribution (OOD) benchmark to assess general model ability. Since the UltraFeedback dataset is not tailored to a specific task, we evaluate models trained with UltraFeedback on this benchmark. Our quality evaluation metrics include: (1) the absolute scores assigned by GPT-4, and (2) the win rate between responses generated by the RLHF baseline and CD-RLHF.

In Table 2, we present the absolute scores assigned by GPT-4, showing that CD-RLHF outperforms the RLHF baseline for all three models. Notably, although the RM scores results of RLHF and CD-RLHF on Llama-3.2-1B and Gemma-7B models are comparable, CD-RLHF demonstrates superior alignment quality when evaluated by GPT-4. Win rates estimated by GPT-4 are depicted in Figure 7, indicate that CD-RLHF achieves win rates of 32.5%, 21.875%, 31%, and 32.5% over RLHF on the Gemma-2B, Gemma-7B, Llama-3.2-1B, and Llama-3.2-3B models, respectively.

Apart from the evaluation of quality, we sample 10 answers for each question to estimate the diversity on this OOD benchmark. Since some of questions belong to the categories which sampling temperatures are set to 0.0, we exclude them in this evaluation. Only reserving the questions from writing, roleplay, stem, and humanities categories, with temperature set to 0.7, 0.7, 0.1, and 0.1, respectively. The results are shown in Table 2, CD-RLHF consistently exhibits improvements on output diversity, both from syntactic and semantic perspectives evaluated by related metrics.

4.6 Extended Experiments: Story Generation

In this section, we assess the output diversity of CD-RLHF in a downstream application: story writing task. We use the validation set of ROC story dataset (Mostafazadeh et al., 2016), covering 1817 stories. The models involved in this test are the Llama-3.2-3B and Gemma-7B trained on the UltraFeedback dataset, trained with RLHF and CD-RLHF. We prompt the model to generate a story with a start sentence provided by the ROC dataset. The generated story is limited to 200 words. We use the configuration `temperature=1.0`, `top_p=0.9`, and `top_k=50`. For evaluation, we use Diveristy, EAD, SelfBLEU, and SentBERT as metrics to assess the output diversity across these stories. The evaluation results are presented in Figure 8. On this creative writing task, CD-RLHF shows more diverse outputs than RLHF evidenced by these metrics from both syntactic and semantic.

5 Conclusion

In this paper, we introduce a novel framework, CD-RLHF, which encourages the policy to explore "novel" context with curiosity, combining with extrinsic reward to optimize both output diversity and alignment quality. Experiments on the TL;DR summarization and Ultrafeedback instruction-following tasks demonstrate CD-RLHF's effectiveness in enhancing output diversity while maintaining alignment quality. Additionally, results on an out-of-distribution benchmark highlight the framework's strong generalization capabilities and the benefits of increased output diversity.

Limitations

While our work demonstrates the benefits of incorporating a curiosity-driven approach into RLHF to enhance output diversity, CD-RLHF has certain limitations. First, we observed that the intrinsic reward scale is significantly larger than the extrinsic reward. To address this, we set the weight parameter η for intrinsic rewards to a small value. However, designing a more suitable ICM that generates intrinsic rewards with a distribution closer to extrinsic rewards could yield better performance. Second, while CD-RLHF mitigates the trade-off between output diversity and alignment quality, this trade-off still persists during the RLHF stage. Finally, although CD-RLHF enhances output diversity without compromising alignment quality, it generally underperforms SFTed models in terms of diversity. In the future, it is promising to further bridge this gap to achieve both the output diversity of SFT and the alignment quality of RLHF.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *CoRR*, abs/2112.00861.

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.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. 2016. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29.

Herbie Bradley, Andrew Dai, Hannah Teufel, Jenny Zhang, Koen Oostermeijer, Marco Bellagente, Jeff Clune, Kenneth Stanley, Grégory Schott, and Joel Lehman. 2023. Quality-diversity through ai feedback. *arXiv preprint arXiv:2310.13032*.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. 2019a. Large-scale study of curiosity-driven learning. In *International Conference on Learning Representations*.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2019b. Exploration by random network distillation. In *International Conference on Learning Representations*.

Louis Castricato, Alexander Havrilla, Shahbuland Matiana, Michael Pieler, Anbang Ye, Ian Yang, Spencer Frazier, and Mark Riedl. 2022. Robust preference learning for storytelling via contrastive reinforcement learning. *arXiv preprint arXiv:2210.07792*.

Yekun Chai, Haoran Sun, Huang Fang, Shuohuan Wang, Yu Sun, and Hua Wu. 2024. Ma-rlhf: Reinforcement learning from human feedback with macro actions. *arXiv preprint arXiv:2410.02743*.

Yekun Chai, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, and Hua Wu. 2023. ERNIE-code: Beyond English-centric cross-lingual pretraining for programming languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10628–10650, Toronto, Canada. Association for Computational Linguistics.

Alex James Chan, Hao Sun, Samuel Holt, and Mihaela van der Schaar. 2024. Dense reward for free in reinforcement learning from human feedback. In *Forty-first International Conference on Machine Learning*.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307.

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. *arXiv preprint arXiv:2310.01377*.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. 2023. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*.

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

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *arXiv e-prints*, pages arXiv–2310.

Nick Haber, Damian Mrowca, Stephanie Wang, Li F Fei-Fei, and Daniel L Yamins. 2018. Learning to play with intrinsically-motivated, self-aware agents. *Advances in neural information processing systems*, 31.

Zhang-Wei Hong, Idan Shenfeld, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James R. Glass, Akash Srivastava, and Pulkit Agrawal. 2024. Curiosity-driven red-teaming for large language models. In *The Twelfth International Conference on Learning Representations*.

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*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Lei Li, Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Ningyu Zhang, and Hua Wu. 2024. Tool-augmented reward modeling. In *The Twelfth International Conference on Learning Representations*.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. 2023. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *Forty-first International Conference on Machine Learning*.

Yiming Liang, Ge Zhang, Xingwei Qu, Tianyu Zheng, Jiawei Guo, Xinrun Du, Zhenzhu Yang, Jiaheng Liu, Chenghua Lin, Lei Ma, et al. 2024. I-sheep: Self-alignment of llm from scratch through an iterative self-enhancement paradigm. *arXiv preprint arXiv:2408.08072*.

Siyang Liu, Sahand Sabour, Yinhe Zheng, Pei Ke, Xiaoyan Zhu, and Minlie Huang. 2022. Rethinking and refining the distinct metric. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. 2024. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. 2017. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2721–2730. PMLR.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 *NeurIPS*.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on*

Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

J Schmidhuber. 1991a. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, pages 222–227.

Jürgen Schmidhuber. 1991b. Curious model-building control systems. In *Proc. international joint conference on neural networks*, pages 1458–1463.

Bradly C Stadie, Sergey Levine, and Pieter Abbeel. 2015. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize from human feedback. *CoRR*, abs/2009.01325.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. 2024. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*.

Fan Wu, Emily Black, and Varun Chandrasekaran. 2024. Generative monoculture in large language models. *arXiv preprint arXiv:2407.02209*.

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. *Advances in Neural Information Processing Systems*, 36:59008–59033.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Forty-first International Conference on Machine Learning*.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. 2024. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*.

Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, et al. 2023. Deepspeed-chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales. *arXiv preprint arXiv:2308.01320*.

Eunseop Yoon, Hee Suk Yoon, SooHwan Eom, Gunsoo Han, Daniel Nam, Daejin Jo, Kyoung-Woon On, Mark Hasegawa-Johnson, Sungwoong Kim, and Chang Yoo. 2024. TLCR: Token-level continuous reward for fine-grained reinforcement learning from human feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14969–14981, Bangkok, Thailand. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *SIGIR*.

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 Experimental Details

A.1 Datasets

TL;DR Summarization In this task, the policy is asked to generate summarizations for Reddit posts. This dataset consists of 93k human-annotated preference pairs and 86k pairs for validation. The trainable pairs are derived from the Reddit TL;DR (Völske et al., 2017) dataset. Additionally, a portion of the validation pairs is sourced from the CNN Daily Mails, which serves as the test set for out-of-distribution generalization.

UltraFeedback Instruction Following In this task, the policy needs to complete the given task, which covers various aspects. This dataset is produced with LLMs assist, the completions are generated with different LLM, and scored by GPT-4 on instruction-following, helpfulness, honesty, and truthfulness aspects. We used the binarized version of UltraFeedback, which use the completion with highest overall score as chosen, and one of the remaining 3 at random as the rejected one. This dataset contains 61.1k instances for training and 2k for evaluation.

We present the data statistics in Table 3.

A.2 Training Details

Following the procedure of previous work (Ouyang et al., 2022), the SFT model and reward model are fine-tuned on the same dataset with RL fine-tuned model to mitigate the distribution gap.

SFT Training We split 20% of the training instances to conduct the supervised fine-tuning. Since the fine-tuning conducted on the preference dataset, we use the prompt and chosen completion as the instruction data. For TL;DR summarization task, we wrap the instruction with the same format as Stiennon et al. (2020): `POST\nSubreddit: r/[SUBREDDIT]\n[PROMPT]\nTL;DR:[CHOSEN]`. For UltraFeedback instruction following task, we use the ChatML format: `<|im_start|>user\n[PROMPT]\n<|im_end|>\n <|im_start|>assistant\n[CHOSEN]\n <|im_end|>\n`.

Reward Modelling We use 40% of the training instances in this stage. The reward model is initialized using the fine-tuned SFT model.

PPO Training The remained 40% of the data is utilized to train the policy model. We initialize the policy and reference model with fine-tuned SFT model, the critic model with reward model. When training Llama-3.2-1B, we use the same reward model as Llama-3.2-3B. We enabled LoRA (Hu et al., 2022) when training 7B model with PPO, and the learning rates of LoRA module in actor and critic model are set to 5e-4, LoRA dimensions are 16. We only insert LoRA into the linear layer of self-attention mechanism.

When implementing the CD-RLHF, the intrinsic reward whitening operation is used to stable the training procedure, which mitigate the scalar gap across samples.

Table 4 lists the hyperparameters used in all stages. The hidden size of encode in ICM is 2 times of hidden size of actor model, and hidden size of ICM is the same as intermediate size of actor model. The experiments are conducted on machines with 8×V100 GPUs or machines with 8×A100 GPUs.

For implementing Sent-Rewards, we adjust the weights of SelfBLEU, SentBERT, and entropy rewards in the reward score. In all experiments, we set the entropy weight to 0.01 and apply the entropy penalty as token-level rewards, similar to KL-divergence. The specific weights for SelfBLEU and SentBERT in each experiment are listed in Table 5.

A.3 Evaluation Details

We evaluate the diversity with the PerInput format. The test set is define as $\mathcal{D}_{\text{test}}$, and $\mathcal{D}_i$ is the i-th subset of $\mathcal{D}_{\text{test}}$, with the completions in $\mathcal{D}_i$ have the same instruction. The details of each metrics are as follows:

$$\text{rep-}n = 100 \times (1.0 - \frac{N_n}{C_n})$$

$$N\text{−gram Distinct}_{\mathcal{D}^i} = \prod_{n=1}^{N} (1.0 - \frac{\text{rep-}n}{100})$$

where N_n is the number of distinct tokens of n-gram, and C_n is the total number of tokens of n-gram.

$$\text{EAD}_{\mathcal{D}^i} = \frac{1}{N} \sum_{n=1}^{N} \frac{N_n}{V[1 - (\frac{V-1}{V})]_n^C}$$

where V is the vocabulary size.

$$\text{SelfBLEU}_{\mathcal{D}^i} = \frac{1}{N} \sum_{n=1}^{N} \text{SelfBLEU}_n(\mathcal{D}^i)$$

Dataset	**Num. of Comparisons**	**Num. of Train Samples**	**Num. of Test Samples**	**Avg. Tokens in Prompt**	**Avg. Tokens in Chosen**	**Avg. Tokens in Rejected**
OpenAI Summarization	179k	92.9k	86.1k	325	35	33
UltraFeedback	63.1k	61.1k	2k	156	282	246

Table 3: Statistics of datasets involved in experiments. The number of tokens are calculated with Gemma-2B tokenizer.

	Hyper-Parameter	Gemma-2B	Gemma-7B	Llama-3.2-1B	Llama-3.2-3B
SFT	Batch size	512 for TL;DR 128 for UltraFeedback	128	256	256
	Epochs	3	1	3	3
	Learning rate	5e-5	2e-5	1e-4	5e-5
	LR scheduler	cosine	cosine	cosine	cosine
	Warmup ratio	0.1	0.1	0.1	0.1
RM	Batch size	64	128	-	64
	Epochs	1	1	-	1
	Learning rate	1e-5 for TL;DR 2e-5 for UltraFeedback	1e-6	-	1e-5
	LR scheduler	cosine	cosine	-	cosine
	Warmup ratio	0.05	0.05	-	0.05
PPO	Batch size	256	256	256	256
	Policy learning rate	8e-6 for TL;DR 1e-5 for UltraFeedback	-	5e-6	5e-6 for TL;DR 8e-6 for UltraFeedback
	Critic learning rate	1e-5	-	1.5e-5	1e-5
	Epochs	1	1	1	1
	PPO epochs	1	1	1	1
	Rollout	1	1	1	1
	Clip ratio	0.2	0.2	0.2	0.2
	λ in GAE	0.95	0.95	0.95	0.95
	γ in GAE	1	1	1	1
	KL coefficient β	0.05	0.05	0.05	0.1
	Max prompt length	512	512	512	512
	Max response length	512	512	512	512
	Warmup ratio	0.1	0.1	0.05	0.05 for TL;DR 0.1 for UltraFeedback
	Temperature	0.8	0.8	0.8	0.8
	Top-p	1.0	1.0	1.0	1.0
	Top-k	50	50	50	50
	η	0.04	0.04	0.06 for TL;DR 0.04 for UltraFeedback	0.08

Table 4: Hyper-parameters for training Gemma-2B, Gemma-7B, Llama-3.2-1B, and Llama-3.2-3B models for all stages.

Model	**SelfBLEU**	**SentBERT**
TL;DR		
Gemma-2B	0.5	0.0
Gemma-7B	0.5	0.5
Llama-3.2-1B	0.5	0.5
Llama-3.2-3B	1.0	1.0
UltraFeedback		
Gemma-2B	1.0	1.0
Gemma-7B	1.0	1.0
Llama-3.2-1B	0.5	0.5
Llama-3.2-3B	0.5	0.5

Table 5: Hyper-parameters for training Sent-Rewards method (Hong et al., 2024) on Gemma-2B, Gemma-7B, Llama-3.2-1B, and Llama-3.2-3B models.

where SelfBLEU_n is the SelfBLEU calculated with n-gram.

$$\text{SentBERT}_{\mathcal{D}^i} = \frac{1}{|\mathcal{D}^i|} \sum_{x_k, x_j \in \mathcal{D}^i} \frac{g(x_k) \cdot g(x_j)}{\|g(x_k)\|_2 \|g(x_j)\|_2}$$

where $g(\cdot)$ is the embedding model.[2]

The final results are averaged across all samples for $\mathcal{D}^i \in \mathcal{D}_{\text{test}}$.

For GPT-4 evaluation, we utilize the `gpt-4o-05-13` to rank the set based on diversity given an instance and two sets of completions, where each set contains 10 completions. We sampled 50 instances from each dataset for evaluation.

For human evaluation, the human annotators are asked to identify which set of completions is more

[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

diverse than the other. Unlike the GPT-4 evaluation, the annotators are given an instance and two sets of completions, each containing 5 completions. We sampled 25 instances from each dataset for evaluation.

B Comparison Evaluation Details

B.1 GPT-4 Evaluation Prompt

For each evaluation, GPT-4 is employed to assess the diversity of two sets of 10 completions generated by different models, serving as a proxy for human evaluation. All experiments use `gpt-4o-05-13`. To ensure fairness, the order of sets is randomized across all experiments.

GPT-4 Evaluation Prompt for Ranking Diversity Score

Given two sets of responses, your task is to identify which set is more diverse compared to the other. The diversity evaluation should assess the variation among the proposed responses. The more similar the responses within a set, the lower the diversity. And also, you need to identify which set of responses is more related to the prompt:
Prompt: `<instances>`
Set 0: `<completion set 1>`
Set 1: `<completion set 2>`
Your identification (starting with reason, and end with your choice, using "MY CHOICE: 0/1/2" to mark the choice, where 0 for Set 0, 1 for Set 1, and 2 for equally diverse):

GPT-4 Evaluation Prompt for Evaluating Alignemnt Quality on TL;DR

You will be given two summaries written for an article. Your task is to pick the better one between them, based on the four criteria. Please make sure you read and understand these instructions carefully.
Relevance - selection of important content from the source. The summary should include only important information from the source document. Annotators were instructed to penalize summaries which contained redundancies and excess information.
Coherence - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby "the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to a coherent body of information about a topic."
Consistency - the factual alignment between the summary and the summarized source. A factually consistent summary contains only statements that are entailed by the source document. Annotators were also asked to penalize summaries that contained hallucinated facts.
Fluency - the quality of the summary in terms of grammar, spelling, punctuation, word choice, and sentence structure.
You should output single character to indicate which summary you think is better. 'A' stands for Summary A and 'B' stands for Summary B. If you think both summaries are equally good, output 'E'.
Article / Post: `<article / post>`
Summary A: `<summary a>`
Summary B: `<summary b>`
Your Choice (only a single character):

GPT-4 Evaluation Prompt for Evaluating Alignemnt Quality on UltraFeedback

You will be given two responses for an instruction written by different assistants. Your task is to pick the better one between them, based on the three aspects: instruction following, helpfulness, and harmlessness. You should output single character to indicate which response you think is better. 'A' stands for Response A and 'B' stands for Response B. If you think both responses are equally good, output 'E'.
Instruction: `<instruction>`
Response A: `<response a>`
Response B: `<response b>`
Your Choice (only a single character):

B.2 Human Evaluation Protocol

Similar to the GPT-4 evaluation, the annotators are given two set of completions generated by two different models, each set contains 5 completions. The human annotators are asked to select the preferred response based on the diversity related criteria:

1. Structure Variety: this considers whether these completions have flexible and diverse structures, adapted to the content.

TL;DR	Win	Tie	Loss
CD-RLHF v.s. SFT	39	0	11
CD-RLHF v.s. RLHF	27	0	23
UltraFeedback	**Win**	**Tie**	**Loss**
CD-RLHF v.s. SFT	31	4	15
CD-RLHF v.s. RLHF	21	7	22

Table 6: GPT-4 pairwise comparisons of alignment quality demonstrate CD-RLHF's superior win rates over both vanilla RLHF and SFT when evaluating the Llama-3.2-3B model on TL;DR and UltraFeedback benchmarks.

2. Paraphrasing Variety: this assesses whether these completions rephrase most of the content in unique and varied ways.

The annotators need to choose the more diverse set based on these criteria.

C Additional Experiments Results

C.1 Alignment Quality Evaluated by GPT-4

To thoroughly assess the alignment quality of models trained with SFT, CD-RLHF, and RLHF, we use GPT-4 to perform pairwise comparisons. Specifically, we employ best-of-N sampling to select the response with the highest reward score, where N is set to 10. This experiment is conducted on the Llama-3.2-3B model using both the TL;DR and UltraFeedback datasets. The results, presented in Table 6, reveal that CD-RLHF maintains alignment quality comparable to RLHF, while significantly outperforming models trained with SFT.

C.2 Experimental Results on Top-k

In Section 3.2.2, the intrinsic reward is valid only when the probability of the selected token lies out of top-k. The k is set to 1 in our experiments. In this section, we set k to 3 and 10 for further analysis. Under these setting, the number of intrinsic rewards decreases with the k increases. These experiments are conducted on TL;DR dataset with Gemma-2B model, and the results are shown in Table 7. We observe that with less intrinsic rewards (higher k), the output diversity deceases while maintaining the alignment quality.

D Algorithm of CD-RLHF

In Algorithm 1, we present the procedure of CD-RLHF using PPO.

E Case Study

To better understand the effectiveness of CD-RLHF and its impact on generated completions, we present examples from the TL;DR summarization task and UltraFeedback instruction-following task, using 2B models. For each case, we display the first 5 completions out of 10.

In Table 8, we show a case involving a student considering pursuing a Master's degree but concerned about financial constraints. In the RLHF completions, keywords like "financially independent", "no debt", "living stipends", and "tuition waivers" appear frequently across responses. CD-RLHF completions, while including these same keywords, also introduce additional details about the author's identity. For instance, response 3 specifies a tuition fee of $40K, and response 4 mentions the author's salary as $23,000 per year and interest about degree. These CD-RLHF responses maintain core information but show greater diversity compared to RLHF.

Table 9 demonstrates a case resembling an instruction generation task, where the model generates questions based on a provided "Background" and "Story". The RLHF completions repeatedly focus on "demand for steel products", "greenhouse gases", and "carbon dioxide", with the term "increase" used consistently. In contrast, CD-RLHF completions exhibit more diversity, with keywords like "global warming", "air quality", "greenhouse gas", and "carbon dioxide". Notably, response 4 uses "decrease" instead of "increase", as allowed by the prompt. Additionally, CD-RLHF completions include an "Explanation" to further elaborate on the solution.

	Diversity↑	EAD↑	SelfBLEU↓	SentBERT↓	RM Score↑
RLHF	0.2132	0.7347	0.3367	0.7024	0.90
CD-RLHF top-1	**0.2839**	**0.7793**	**0.2590**	**0.6720**	**0.95**
CD-RLHF top-3	0.2690	0.7747	0.2669	0.6771	0.94
CD-RLHF top-10	0.2613	0.7670	0.2834	0.6833	**0.95**

Table 7: Results on TL;DR dataset with different top-k settings for CD-RLHF.

Algorithm 1 Framework of Curiosity-Driven RLHF.

Require: the train dataset $\mathcal{D} = \{x_0, x_1, \ldots, x_n\}$
Require: Policy model: π_{policy};Reference model: π_{ref}; Critic model: π_{critic}; Reward model: π_{rm}; ICM model: π_{ICM}
for x in $\mathcal{D}$ **do**
    Make experience using policy model $y := \pi_{\text{policy}}(x)$;
    Get value $V(s_t) := \pi_{\text{critic}}(x, s_t)$ at every time step $t \in [0, |y|)$;
    Get reward score at current experience $R := \pi_{\text{rm}}(x, y)$;
    Compute KL as token-level reward $-\beta D_{\text{KL}}(\pi_{\text{policy}}(a_t) \| \pi_{\text{ref}}(a_t))$
    Obtain extrinsic rewards $r^{(e)} = R - \beta D_{\text{KL}}(\pi_{\text{policy}}(\cdot) \| \pi_{\text{ref}}(\cdot))$
    Compute intrinsic rewards $r^{(i)}$ following demonstration in Section 3.2.2
    Combine intrinsic and extrinsic rewards: $r = r^{(e)} + \eta \cdot r^{(i)}$
    Obtain advantage function A_t and Q-value function Q_t with GAE$(V(s_t), r)$
    Optimize policy with $\mathcal{L}_{\text{ppo}} = \hat{\mathbb{E}}\left[\min\left(\frac{\pi_{\text{policy}}(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)} A_t, \text{clip}(\frac{\pi_{\text{policy}}(a_t|s_t)}{\pi_{\text{ref}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon) A_t\right)\right]$
    Optimize critic with $\mathcal{L}_{\text{critic}} = \hat{\mathbb{E}}\left[\|V(s_t) - Q_t\|^2\right]$
    Optimize ICM with $\mathcal{L}_{\text{ICM}} = \frac{1}{2}\|\hat{\phi}(s_{t+1}) - \phi(s_{t+1})\|_2^2$
end for

Prompt	POST Subreddit: r/personalfinance **(Let me know if there's a better subreddit for this or if it has been profusely addressed before. I have been browsing this subreddit for a while and have not come across it.)** I am 23, recently financially independent, with a degree in Linguistics that by itself doesn't do much, but coupled with a Master's could be quite useful. However, getting my masters would put me about $40,000 in the hole. I just started off, am NOT currently in debt, and don't really want to start my financially independent life by going in debt. When I went to undergrad, my dad paid for everything and I didn't have to worry about financial aid, so I have no idea how it works and have never dealt with the system. My question(s) is/are: 1. I know some people get living stipends and tuition wavers for graduate school. How common is this? Have you done so before? 2. Would you recommend going to graduate school? Right now I am making $23,000 a year and with a degree in speech language pathology could be making significantly more (and it's a field I'm interested in.) Basically (TL;DR:
	RLHF
Resp. 1.	23 year old recently financially independent, wondering how to pay for graduate school, if living stipends and tuition wavers are common and if it would be worth putting myself in debt.
Resp. 2.	23 year old recently financially independent with no debt looking for 1) help with financial aid/scholarships for graduate school as tuition/housing are prohibitively high and 2) advice on if it would be beneficial to go back to school.
Resp. 3.	23 year old recently financially independent, looking at the possibility of going to graduate school to build skills and career development. Know some people get living stipends and tuition wavers for graduate school, how common is that and would you recommend going?
Resp. 4.	23, recently financially independent with no debt, wanting to know more about living stipends and tuition wavers for graduate school. Would it be worthwhile to attend?
Resp. 5.	23 year old recently financially independent with a degree but no clue how to proceed with grad school. Know some people get living stipends and tuition wavers; would you recommend going to graduate school?
	CD-RLHF
Resp. 1.	23, recently financially independent, wanting to apply for graduate school next fall (masters), seeking advice on getting living stipends and tuition wavers. Would it be worth pursuing?
Resp. 2.	23/f recently financially independent, recently thinking about pursuing grad school to further my career. Do you get living stipends and tuition wavers for grearndschools and would you recommend going?
Resp. 3.	23, recently financially independent, looking at grad school for 40K tuition in field that could earn much more monetarily. Want to know if it is possible/worth it for grad school as well as advice on grad school process.
Resp. 4.	23 year old recently financially independent - recently learned about graduate school. Wondering whether anyone has been granted living stipends or tuition waivers - would you recommend going to grad school? Right now making $23,000 a year and interested in SPEECH/PHARM/PEDIATRIC psychology.
Resp. 5.	23 yr old college alumni with degree in linguistics needing financial assistance for graduate school-has anyone gotten living stipends and/or tuition wavers for graduate school? Would it be a wise move financially and is there any info on the process/realities?

Table 8: Example responses from TL;DR dataset. We mark the keywords that showed the first time with green, and red for repeated keywords for better illustration.

Prompt	You are given a background paragraph that describes one or more causal or qualitative relationships such as a relationship in economics or a scientific law and a story that makes use of the concepts or the relationship described in the provided paragraph. You need to come up with a question about the story that requires an understanding of the relationship described in the background paragraph. The generated question should not be answerable without both the background and story. Write a question about the story that requires a relationship in the background paragraph to answer. Check whether your question requires both the background paragraph and the story to answer. If your question can be answered by only one of these, try to rewrite your question so that it requires both. Note that "flipping" a word in the question can give rise to a new question where the answer will be different from the original question. Often, "flipping" a word gives rise to flipping the direction of the relationship, for example, from "increase" to "decrease." One example: Story: The AAA auto club does a weekly report on the prices of gas and diesel in various cities. Historically, the prices have be fairly stagnant, but this week their report garnered widespread attention as the prices of gas and diesel plunged from last week in Seattle by \$0.40 each. However, just across the border in Vancouver, the cost of gas went up dramatically by \$0.50. Solution is here: Which city will have an increase in demand for gas? Explanation: This is a good question as it requires understanding of the relationship described in both the background and story (the background mentions "a fall in price will increase the quantity demanded" and the story cites the decreasing price of gas in Seattle). Now, solve this: Background Paragraph: Pollutants also affect the atmosphere through their contribution to global warming. Global warming is an increase in the Earth's temperature. It is thought to be caused mostly by the increase of greenhouse gases like carbon dioxide. Greenhouse gases can be released by factories that burn fossil fuels. Over the past 20 years, burning fossil fuels has produced about three-quarters of the carbon dioxide from human activity. The rest of the carbon dioxide in the atmosphere is there because of deforestation, or cutting down trees (Figure below). Trees absorb carbon dioxide during cellular respiration, so when trees are cut down, they cannot remove carbon dioxide from the air. Story: Tokyo recently had a large explosion in one of their largest steel manufacturing plants. The combination of heat, steel, and manufacturing materials released an extraordinary amount of carbon dioxide into the atmosphere. Meanwhile, Osaka has dramatically decreased the amount of carbon dioxide it releases through the installation of many public transportation projects and tree planting initiatives. Solution:
	RLHF
Resp. 1.	Which city will have an increase in demand for steel products?
Resp. 2.	Which city will experience an increase in demand for steel products due to the Tokyo explosion?
Resp. 3.	Which city will experience an increase in the release of greenhouse gases following the Tokyo explosion?
Resp. 4.	Which city will have an increase in the amount of carbon dioxide released into the atmosphere?
Resp. 5.	Which city experienced an increase in the release of carbon dioxide following the Tokyo explosion?
	CD-RLHF
Resp. 1.	Which city experienced an increase in global warming as a result of the Tokyo steel plant explosion? Explanation: ...
Resp. 2.	Which city experienced an increase in air quality due to the decrease in carbon dioxide emissions? Explanation: This question requires understanding of the relationship described in the background paragraph (global warming caused by greenhouse gases) and the story (the accidental explosion in Tokyo released carbon dioxide into the atmosphere, while Osaka decreased their carbon dioxide emissions through public transportation and tree planting initiatives).
Resp. 3.	Which city's greenhouse gas emission will increase as a consequence of the Tokyo explosion? Explanation: ...
Resp. 4.	Which city will have a decrease in demand for steel products? Explanation: In the given background, it is stated that the excessive release of carbon dioxide from a large-scale manufacturing plant in Tokyo can lead to a decrease in demand for steel products. On the other hand, Osaka has taken initiatives to decrease carbon dioxide emissions through public transportation and tree planting projects, which can potentially increase the demand for steel products. Therefore, this question requires an understanding of the relationship between carbon dioxide emissions and demand for steel products, as well as the context of the given stories.
Resp. 5.	Which city will have an increase in carbon dioxide emissions? Explanation: ...

Table 9: Example responses from Ultrafeedback dataset. We omit some explanations in responses. We mark the keywords that showed the first time with green, and red for repeated keywords for better illustration.