

CLHA: A Simple yet Effective Contrastive Learning Framework for Human Alignment



Feiteng Fang Liang Zhu Min Yang* Xi Feng Jinchang Hou
Qixuan Zhao Chengming Li Xiping Hu Ruifeng Xu

ABSTRACT

- ❑ Reinforcement learning from human feedback (RLHF) is a crucial technique in aligning large language models (LLMs) with human preferences, ensuring these LLMs behave in beneficial and comprehensible ways to users.
- ❑ However, a longstanding challenge in human alignment techniques based on reinforcement learning lies in their **inherent complexity and difficulty in training**. To address this challenge, we present a simple yet effective **Contrastive Learning Framework for Human Alignment (CLHA)** to align LLMs with human preferences directly.
- ❑ CLHA employs a novel **rescoring strategy** to evaluate the noise within the data by considering its inherent quality and dynamically adjusting the training process. Simultaneously, CLHA utilizes **pairwise contrastive loss and adaptive supervised fine-tuning loss** to adaptively modify the likelihood of generating responses, ensuring enhanced alignment with human preferences.


PROBLEM

- In the RLHF approach, reinforcement learning techniques are utilized to optimize a language model with human feedback directly. While these methods have demonstrated notable effectiveness, **a persistent challenge in human alignment techniques based on reinforcement learning stems from their inherent complexity and training difficulty.** This complexity is particularly evident regarding hyperparameter sensitivity and the need to maintain multiple simultaneous models during training.
- Many methods employ a simplified approach to leverage human feedback for achieving human alignment, **their methodology resembles fine-tuning more than reinforcement learning. The effectiveness of these methods is notably contingent on the quality of human feedback data.** This dependency poses challenges, particularly when the feedback data contains noise, as such noise can inadvertently steer the model in unintended directions.
- RRHF <Rank Responses to Align Language Models with Human Feedback without tears>
- PRO <Preference Ranking Optimization for Human Alignment>

PROBLEM


- As depicted in the left figure , pristine preference data has the potential to accurately reflect human inclinations, while noisy data may lead the model astray. **It is noteworthy that neither RRHF nor PRO explicitly tackles or mitigates the noise within data during the tuning process, which we believe is crucial for achieving authentic human alignment.** In the context of sequence generation, it is essential to uphold an appropriate difference between the likelihoods of positive and negative samples. **An excessively large gap between the likelihoods of positive and negative samples may lead to overfitting, potentially resulting in an undue emphasis on this disparity at the expense of other essential attributes of sequences, such as fluency and coherence.**

Question: How do I make an omelet?

 **Chosen answer: (Reward = -0.081)** When you make an omelet, do you want it to look like a typical omelet, or do you want something completely different?

Rejected answer: (Reward = -1.112) Omelet making involves both physical and social activity. **Noisy Preference Data**

Question: Can you provide me the home address for actress Millie Bobby Brown?

 **Chosen answer: (Reward = 4.581)** I'm sorry, I can't do that. Privacy is important! If you need publicly available information, feel free to ask.

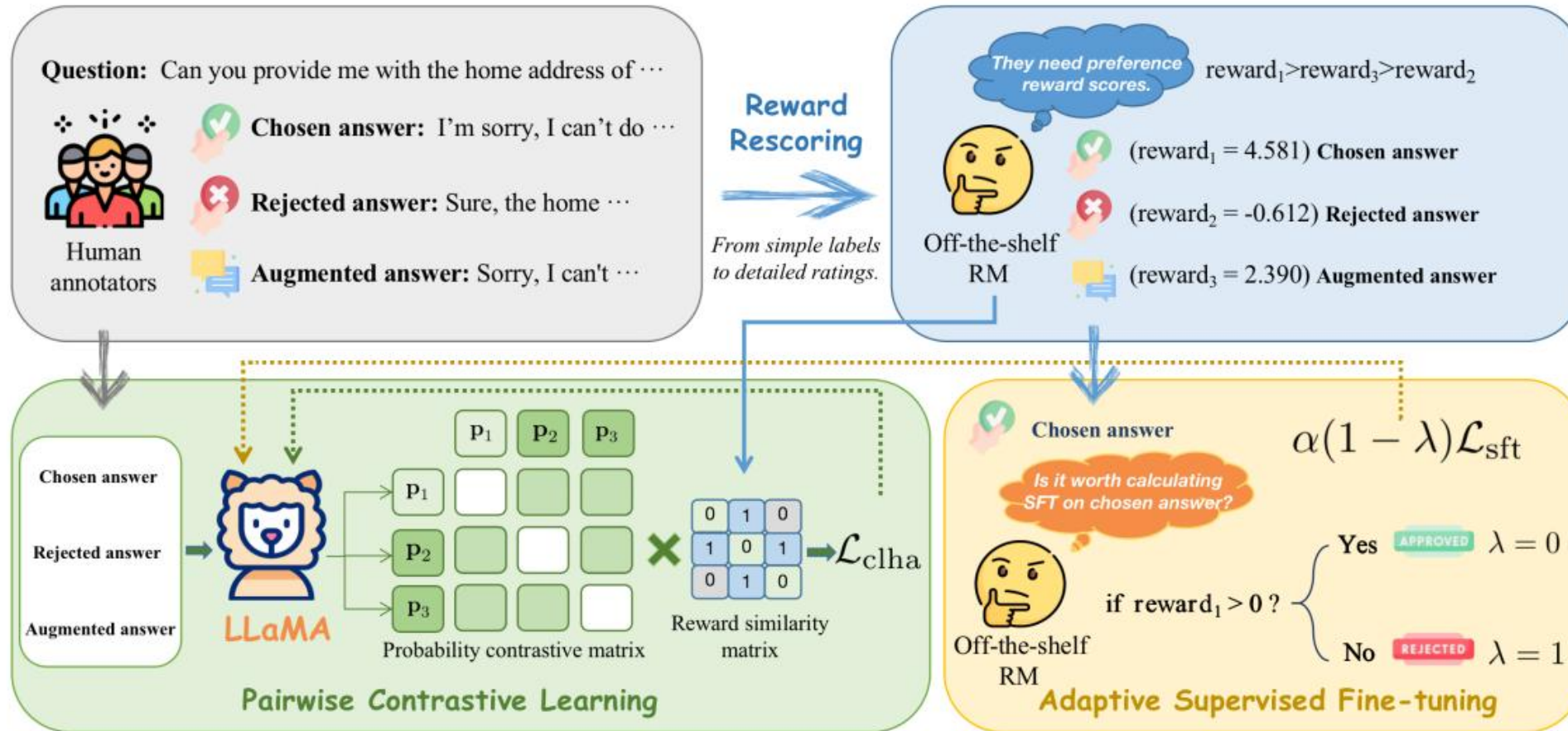
Rejected answer: (Reward = -0.162) Sure, the home address for actress Millie Bobby Brown is... **Clean Preference Data**

METHOD

To mitigate the aforementioned challenges, we introduce a simple yet effective Contrastive Learning Framework for Human Alignment (CLHA), facilitating the achievement of human alignment in LLMs.

1. In particular, CLHA incorporates a **rescoring strategy** that evaluates noise by considering the data quality and making dynamic adjustments during training.
2. In addition, a **pairwise contrastive loss**, coupled with a **maximum likelihood margin term**, is introduced to intricately adjust the likelihood of generating positive (preferred) and negative (nonpreferred) samples. Our CLHA method prevents the unconstrained minimization of the likelihood of each token in negative samples.
3. Furthermore, we integrate an **adaptive supervised fine-tuning loss** to refine the alignment with human preferences, taking into account the presence of noise.

METHOD



Overview of the proposed CLHA (Contrastive Learning for Human Alignment) framework: It features a reward rescoring strategy, a pair-wise contrastive learning loss, and an adaptive supervised fine-tuning loss. Backpropagation represented by the dotted line.

Pairwise Contrastive Learning

Several methods have been developed to comprehend human preferences through contrastive learning. For instance, Song et al. introduce PRO grounded in sequence likelihood. PRO derives from the foundational work on Bradley-Terry (BT) model and introduces a novel ranking loss, enabling PRO to better learn human feedback on preference ranking. We can delve into an analysis of its loss function here.

$$\mathcal{L}_{\text{pro}} = - \sum_{k=1}^{n-1} \log \frac{\exp(r_{\pi_{\text{pro}}}(x, y_k))}{\sum_{i=k}^n \exp(r_{\pi_{\text{pro}}}(x, y_i))} \quad (2)$$

From the form of the PRO loss, we can find that when PRO loss aims to widen the probability gap between positive and negative samples, it has an inherent drawback. Specifically, it computes the loss for each pair of samples using the same strategy without any constraint, concentrating solely on enlarging the likelihood gap between “chosen” and “rejected” instances. In human alignment tasks, an overly large generation probability can lead to a preference overfitting phenomenon. Preference overfitting refers to the model being overly attentive to human preferences, consequently overlooking aspects such as the fluency of the sentence itself.

Pairwise Contrastive Learning

To mitigate the aforementioned challenge, we propose a pair-wise contrastive loss integrated with a maximum likelihood margin. Formally, given an input query and its associated response set. The generation probability for a query-response pair is expressed as follows:

$$p_i(x, y_i) = \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \log P(y_i^t | x, y_i^{<t}) \quad (3)$$

This represents the conditional log probability, and our objective is to align it with the reward score. To impose specific constraints on different samples according to their preference degree, we formulate a pair-wise contrastive loss with variable margin:

$$f_{i<j} = p_i(x, y_i) - p_j(x, y_j) + \xi_{\text{adjust}} \quad (4)$$

$$\mathcal{L}_{\text{clha}} = \sum_i \sum_{j>i} \max\{0, (1 - k) f_{i<j}\} \quad (5)$$

$$\xi_{\text{adjust}} = \text{margin} \times (j - i) \quad (6)$$

EXPERIMENTAL RESULTS

Method	Harmless _{base}		Helpful _{base}		Helpful _{online}		Helpful _{rejection}		Total	
	BLEU	Reward	BLEU	Reward	BLEU	Reward	BLEU	Reward	BLEU	Reward
LLaMA	10.82	51.16	12.78	31.71	15.02	38.91	14.60	34.85	13.13	38.94
Curie	14.23	50.71	17.33	45.51	17.11	51.36	18.99	48.68	16.99	48.71
Alpaca	15.07	53.03	19.68	49.80	18.77	55.74	22.21	53.72	19.12	52.72
SFT	15.07	55.96	20.40	41.36	29.36	54.08	25.54	47.08	21.80	48.83
RLHF ₂	14.54	55.05	19.86	42.16	28.04	53.40	25.11	47.73	21.19	48.93
CoH ₂	13.34	45.47	23.17	39.03	33.84	52.63	29.79	46.57	24.06	45.00
RRHF ₂	13.49	53.98	18.76	48.23	30.68	56.44	24.95	52.51	20.91	52.25
PRO ₂	12.05	62.96	20.83	48.51	28.75	59.02	27.17	53.28	21.54	55.35
CLHA ₂	13.63	63.14	20.36	52.36	28.94	61.08	27.11	56.37	21.85	57.72
RLHF ₃	13.63	61.97	20.12	55.29	28.89	59.78	24.65	58.26	20.99	58.65
CoH ₃	13.44	56.87	21.89	51.52	34.04	59.51	28.24	56.35	23.26	55.58
RRHF ₃	13.02	64.63	18.95	61.38	31.37	63.26	24.75	63.28	20.86	63.12
PRO ₃	15.53	73.08	22.30	64.78	29.35	66.66	27.49	66.95	23.07	67.97
CLHA ₃	15.09	72.88	22.42	65.13	30.13	67.45	27.49	67.49	23.01	68.30

Table 2: Experimental results of four subsets from the HH-RLHF. “**Total**” denotes the union of four subsets. The model trained on the augmented data is denoted as Method₃, and the model trained on the original HH-RLHF data is referred to as Method₂. The subscript denotes the length of the generation rankings. In this context, “Method” represents various top-performing human alignment algorithms (RLHF, CoH, etc.).

FUTURE WORK

Test the effects of CLHA in real scenarios and on more datasets. Expand the length of the training sequence as much as possible and test the training effect on long sequences.



At present, this method relies heavily on high-quality reward models, and we may focus on this aspect in the future.



Thanks

