Research Preparation Criterion

Student

Shuyue Jia (U62343813)

Ph.D. Student in Computer Engineering

Supervisor

Dr. Vijaya B. Kolachalama

Report Title

Preference Alignment via Reinforcement Learning from Human Feedback

Selected Paper

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano,

and G. Irving, "Fine-tuning language models from human preferences,"

arXiv preprint arXiv:1909.08593, Sept. 2019.

Presentation

May 14$^{\text{th}}$, 2024

Abstract

Autoregressive Large Language Models (LLMs) demonstrate remarkable capabilities across various domains, sparking the emergence of Artificial General Intelligence (AGI). By predicting subsequent tokens, these models harness extensive, polydisciplinary knowledge and elicit reasoning abilities. However, although language models are becoming more advanced, ensuring their alignment with human values and ethics remains a challenge. To address this issue, Reinforcement Learning from Human Feedback (RLHF) has gained traction as a method to train language models based on human preferences. Drawing inspiration from preference models such as the Bradley-Terry Model, we propose incorporating reward learning into the RLHF framework for preference alignment. Our extensive experiments indicate that language models tend to align with human values through reward maximization under a Kullback-Leibler (KL) divergence constraint. In the future, we plan to explore the potential of using direct policy optimization techniques for analyzing model reliability and robustness, such as the consistency of Artificial Intelligence (AI) systems.

Keywords: Large Language Models (LLMs), Reinforcement Learning from Human Feedback (RLHF), Reward Learning, Preference Models, Alignment.

1 Introduction

Recent advancements in Large Language Models (LLMs) have propelled the field of Artificial General Intelligence (AGI) forward, showcasing emergence and homogenization capabilities [1]. These models, by predicting subsequent tokens, exhibit the ability to leverage vast interdisciplinary knowledge and foster reasoning. However, as LLMs grow in sophistication, the challenge of ensuring their alignment with human values and ethics becomes increasingly evident [2]. In this context, understanding how language models can be trained to better reflect human values is crucial for the responsible advancement of Artificial Intelligence (AI) technologies. This paper aims to contribute to this understanding by proposing and evaluating methods for preference alignment in LLMs.

Addressing the challenge of aligning language models with human values and preferences has become increasingly crucial. In response, Reinforcement Learning from Human Feedback (RLHF) has emerged as a promising approach for training language models [2–4]. RLHF represents a significant shift in the paradigm of model training, where the model learns directly from human feedback. This feedback is essential for guiding the model toward generating outputs that are more aligned with human preferences.

Within the RLHF framework, existing literature can be broadly categorized into two types of methods: reward-based and reward-free approaches. Reward-based methods leverage preference models to train a reward predictor from human preference data [5]. This learned reward model is then used to incentivize desired behaviors during Reinforcement Learning (RL) [2–4,6]. Reward-free methods, on the other hand, directly optimize language models by satisfying preference signals without the need for explicitly training a reward model [7–10]. This approach circumvents potential issues such as reward model overoptimization and the complexity of multi-stage training processes [11]. One-stage reward-free methods appear promising, but they may result in a biased distribution that favors unseen responses, ultimately declining the quality of the learned policy [12]. This paper will concentrate on using reward-based RLHF to align human preferences.

2 Problem Formulation

RLHF aims to ensure that AI models behave in accordance with human preferences and values. Thus, our goal is to fine-tune a language model, also known as policy π_θ, which is parameterized by θ, to generate responses that are preferred by humans. When providing a prompt $\mathbf{x}$, the language model π_θ generates a response $\mathbf{y}$ by predicting the next token.

$$\pi_\theta(\mathbf{y} \mid \mathbf{x}) = \prod_t \pi_\theta \left(y_t \mid \mathbf{x}, \mathbf{y}_{<t}\right), \tag{1}$$

where y_t represents the t^{th} token in the response $\mathbf{y}$, and $\mathbf{y}_{<t}$ represents the tokens generated before the current y_t token.

The Trust Region Policy Optimization (TRPO) algorithm is widely used for optimizing policies [13].

$$\begin{aligned} \underset{\theta}{\text{maximize}} \quad & \mathbb{E}\left[\frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\theta_{\text{ref}}}(\mathbf{y}|\mathbf{x})} A\right], \\ \text{s.t.} \quad & \mathbb{E}\left[\mathrm{KL}\left[\pi_{\theta_{\text{ref}}}(\cdot \mid \mathbf{x}), \pi_\theta(\cdot \mid \mathbf{x})\right]\right] \leq \delta. \end{aligned} \tag{2}$$

The reference model, denoted by $\pi_{\theta_{\text{ref}}}$, undergoes supervised fine-tuning in the initial stage. A estimates the advantage function, which represents the reward model in our case, and $\mathrm{KL}(\cdot)$ is the Kullback-Leibler divergence. The key idea behind TRPO is to transform the original constrained

optimization problem into an unconstrained problem by introducing a penalty term, *i.e.*, the method of Lagrange multipliers, to the objective function [13, 14]. The objective function is maximized by optimizing the policy π_θ in order to solve the problem of RLHF.

$$\underset{\theta}{\text{maximize}}\ \mathbb{E}\left[\frac{\pi_\theta\left(\mathbf{y}\mid\mathbf{x}\right)}{\pi_{\theta_{\text{ref}}}\left(\mathbf{y}\mid\mathbf{x}\right)}A-\beta\text{KL}\left[\pi_{\theta_{\text{ref}}}\left(\cdot\mid\mathbf{x}\right),\pi_\theta\left(\cdot\mid\mathbf{x}\right)\right]\right], \tag{3}$$

where β is an adaptive coefficient that controls the penalty degree. The penalty term, also known as entropy bonus, is of great importance as it prevents the policy π_θ from deviating too far away from the original reference $\pi_{\theta_{\text{ref}}}$.

3 Proposed Method

As shown in Figure 1, there are three steps for RLHF: supervised fine-tuning (SFT), reward learning, and policy optimization. The SFT significantly improves the models' abilities across different domains. The reward model is trained on human preference data and can rank preferences. The language model is ultimately fine-tuned to enhance its performance and maximize the overall reward. During the process of collecting more preference data, the models' responses are aligned with human preferences and values through a sequence of reward learning and policy optimization. We will now discuss each part in detail.

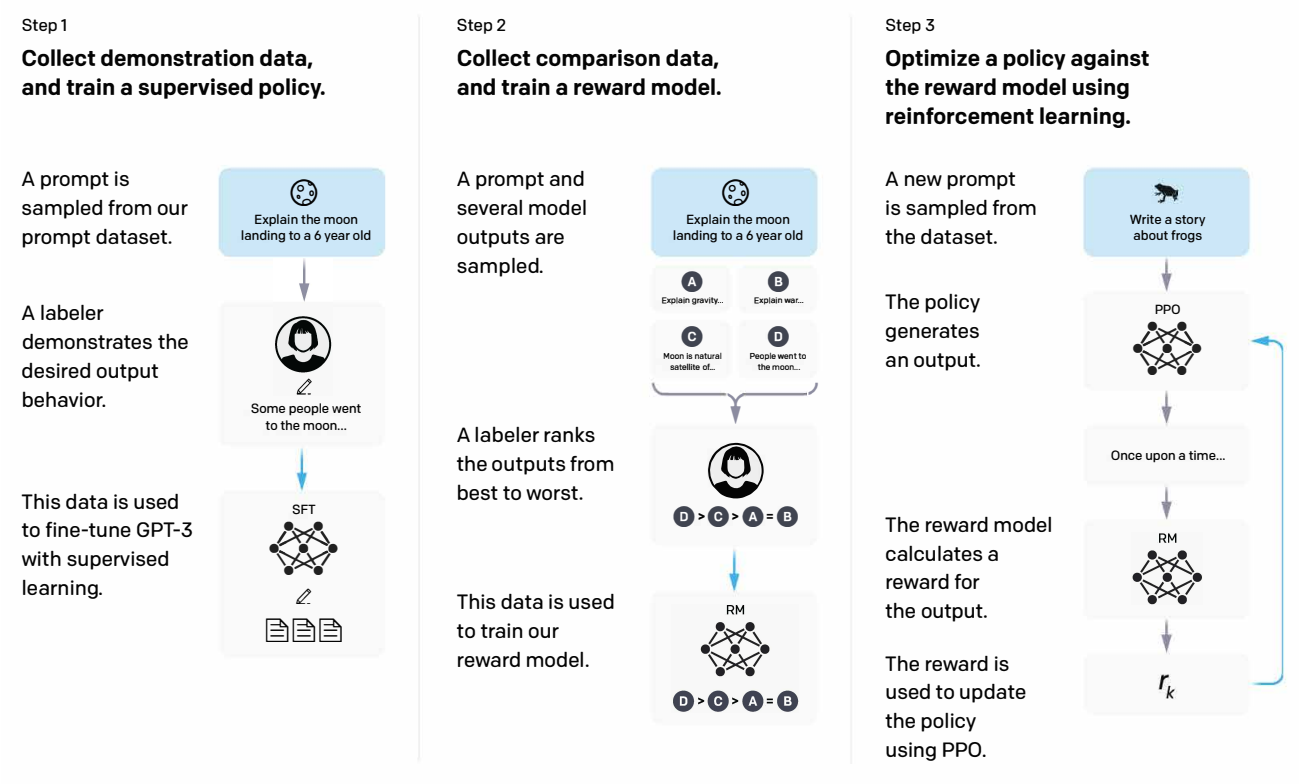

Figure 1: The three-step diagram of the RLHF. Image by courtesy of Ouyang *et al.* [4].

3.1 Supervised Fine-tuning

The first stage is SFT. The language model π_θ is fine-tuned on a diverse and high-quality dataset $\mathcal{D}_{\text{SFT}}$ consisting prompt $\mathbf{x}$ and ground truth response $\mathbf{y}$. This step is well-known as instruction tuning. The goal is to maximize the log-likelihood of response $\mathbf{y}$, conditioned on the prompt $\mathbf{x}$.

$$\mathcal{L}_{\pi_\theta}(\theta) = \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_{\mathrm{SFT}}} \log P(\mathbf{y} \mid \mathbf{x}). \tag{4}$$

3.2 Reward Learning

Based on the above TRPO framework, a reward model r_ϕ is introduced for human preferences representation. Typically, the Bradley-Terry Model is employed to connect rewards with human rating preferences, which estimates the probability of the preferred response $\mathbf{y}_w$ being better than the dispreferred response $\mathbf{y}_l$ [5].

$$\begin{aligned} P\left(\mathbf{y}_w \succ \mathbf{y}_l \mid \mathbf{x}\right) &= \frac{\exp\left(r_\phi\left(\mathbf{x}, \mathbf{y}_w\right)\right)}{\exp\left(r_\phi\left(\mathbf{x}, \mathbf{y}_w\right)\right) + \exp\left(r_\phi\left(\mathbf{x}, \mathbf{y}_l\right)\right)} \\ &= \sigma\left(r_\phi\left(\mathbf{x}, \mathbf{y}_w\right) - r_\phi\left(\mathbf{x}, \mathbf{y}_l\right)\right), \end{aligned} \tag{5}$$

in which σ is a logistics function, *e.g.*, sigmoid function.

Given a collection of preference data $\mathcal{D} = \{(\mathbf{x}, \mathbf{y}_w, \mathbf{y}_l)\}$, the reward model is learned by minimizing the binary ranking loss, *i.e.*, the negative log-likelihood, as follows.

$$\mathcal{L}_{r_\phi}\left(\phi\right) = -\mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}}\left[\log\sigma\left(r_\phi\left(\mathbf{x}, \mathbf{y}_w\right) - r_\phi\left(\mathbf{x}, \mathbf{y}_l\right)\right)\right]. \tag{6}$$

In practice, the reward model r_ϕ is initialized from either a reference model $\pi_{\theta_{\mathrm{ref}}}$ or a reference model with a random output linear layer.

3.3 Policy Optimization

According to Equation 3, our RLHF goal is to maximize the reward:

$$\mathcal{L}_{\pi_\theta}\left(\theta\right) = \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}\ ,\mathbf{y}\sim\pi_\theta}\left[r_\phi(\mathbf{x}, \mathbf{y}) - \beta\log\frac{\pi_\theta(\mathbf{y} \mid \mathbf{x})}{\pi_{\mathrm{ref}}\left(\mathbf{y} \mid \mathbf{x}\right)}\right]. \tag{7}$$

The reward model r_ϕ takes the prompt $\mathbf{x}$ and response $\mathbf{y}$ as input and outputs a scalar value.

4 Discussions

Four Natural Language Processing (NLP) tasks are used to extensively validate the effectiveness of the proposed method: continuation text tasks with positive sentiment and physically descriptive language, as well as summarization tasks on the TL;DR and CNN/Daily Mail datasets. It has been observed that in mock sentiment tasks, it is possible to achieve superior performance with only a small amount of human-labeled data. Besides, although the model primarily relies on selective copying to summarize, it manages to maintain response truthfulness. Let's now consider the benefits and drawbacks of this approach.

Built upon the TRPO framework, there is a theoretical guarantee that the policy optimization is converging [13]. Furthermore, its practical effectiveness is another significant advantage. The state-of-the-art AI models, such as GPT-4 and Claude 3, are fine-tuned using this method for following instructions and aligning with human preferences [1, 2, 4, 14]. Additionally, separately optimizing the reward model and policy improves interpretability, encouraging interchangeable updates in an online fashion. In this way, updating information and human preferences in real time would pave the road for AGI. Last but not least, RLHF improves the accuracy of machine learning (ML) models and enhances user satisfaction. Incorporating more human feedback in

the loop noticeably improves the model's performance.

Nevertheless, as shown in Section 3, the proposed algorithm involves three separate steps: SFT, reward learning, and policy optimization. Such a multi-stage framework, although effective, leads to a time-consuming and laborious training process [7]. Besides, policy optimization is sensitive to hyperparameters and can result in an unstable training process [7]. Applying different sets of hyperparameters could achieve local minimum coverage, leading to unsatisfying outcomes. Furthermore, reference [12] highlights a reward misspecification issue due to narrow distribution coverage on the human preference dataset. It may produce unpredictable results when presented with out-of-distribution (OOD) data. Finally, this method is mostly applied to learn separate rewards, each with a specific optimization goal. It would further complicate the process of training the model. The issue of how to conduct reward learning when dealing with multiple objectives and complex goals remains unresolved.

5 Future Work

In the future, policy optimization without explicitly training a reward model is expected to become more popular due to its simplicity. Recently, there has been an increase in the number of related works in the field [8–10]. To be more specific, the policy can be directly optimized to best satisfy human preferences. For instance, reference [7] implicitly parameterizes a reward model through the theoretically proven relationship between the reward function and the optimal policy. In this way, the policy can be directly optimized through human preference or ranking data, eliminating the need to model rewards. As a result, it significantly reduces the laborious process of data collection and reward model training.

Additionally, RLHF can be employed to enhance model reliability and robustness by specifically enforcing consistency. By incorporating consistency-related ranking data, the model's reliability can be improved through the RLHF process. In specific, it optimizes the policy from human preferences to generate more consistent responses with semantics-preserving prompting. However, effectively collecting human-preferred consistency ranking data is a major challenge behind this idea. Recently, we have proposed a benchmark database, *GSM8K-Consistency*, for analyzing the consistency of arithmetic reasoning on GSM8K, a math problem semantics-preserving perturbation benchmark [15, 16] [1]. We believe it can be helpful for evaluating the consistency of the arithmetic reasoning capability of LLMs [17]. Besides, we have developed *PromptCraft*, a toolkit for prompt robustness analysis, with perturbation at the character, word, and sentence levels, respectively [18] [2]. After benchmarking and toolkit construction, we plan to explore the potential applications of RLHF to improve model reliability and robustness.

Finally, dealing with the challenge of distribution shift when handling OOD samples remains an open question in this field. According to reference [12], the learned reward model may assign a higher probability to OOD data because of the narrow distribution coverage of the preference dataset. As a consequence, it is crucial to develop novel approaches and theoretical foundations to mitigate reward misspecification in order to improve the robustness of advanced AI systems.

[1] The benchmark database is publicly available at https://huggingface.co/datasets/shuyuej/GSM8K-Consistency for training and performance evaluation.

[2] The built toolkit can be accessed at https://github.com/SuperBruceJia/promptcraft and https://pypi.org/project/promptcraft for future scientific research.

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, Aug. 2021.

[2] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, Sept. 2019.

[3] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 3008–3021, Dec. 2020.

[4] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 27730–27744, Nov. 2022.

[5] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. The method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, Dec. 1952.

[6] L. Gao, J. Schulman, and J. Hilton, "Scaling laws for reward model overoptimization," in *International Conference on Machine Learning (ICML)*, pp. 10835–10866, PMLR, July 2023.

[7] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, Dec. 2023.

[8] Z. Yuan, H. Yuan, C. Tan, W. Wang, S. Huang, and F. Huang, "RRHF: Rank responses to align language models with human feedback without tears," *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2023.

[9] T. Liu, Y. Zhao, R. Joshi, M. Khalman, M. Saleh, P. J. Liu, and J. Liu, "Statistical rejection sampling improves preference optimization," *International Conference on Learning Representations (ICLR)*, May 2024.

[10] F. Song, B. Yu, M. Li, H. Yu, F. Huang, Y. Li, and H. Wang, "Preference ranking optimization for human alignment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 18990–18998, Mar. 2024.

[11] A. Ramé, N. Vieillard, L. Hussenot, R. Dadashi, G. Cideron, O. Bachem, and J. Ferret, "WARM: On the benefits of weight averaged reward models," *arXiv preprint arXiv:2401.12187*, Jan. 2024.

[12] S. Xu, W. Fu, J. Gao, W. Ye, W. Liu, Z. Mei, G. Wang, C. Yu, and Y. Wu, "Is DPO superior to PPO for LLM alignment? A comprehensive study," *arXiv preprint arXiv:2404.10719*, Apr. 2024.

[13] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning (ICML)*, pp. 1889–1897, PMLR, July 2015.

[14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, July 2017.

[15] S. Jia, "GSM8K-Consistency Benchmark." `https://huggingface.co/datasets/shuyuej/GSM8K-Consistency`, 2023.

[16] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, Oct. 2021.

[17] S. Jia, "Awesome LLM Self-Consistency." `https://github.com/SuperBruceJia/Awesome-LLM-Self-Consistency`, 2023.

[18] S. Jia, "PromptCraft: A prompt perturbation toolkit." `https://github.com/SuperBruceJia/promptcraft`, 2023.