International Journal of Research and Review in Applied Science, Humanities, and Technology
Vol 3, Issue 2 April-June 2026 ISSN: 3048-975X
https://ijrasht.com/

Learnable Reward Weighting in Multimodal RLHF: A Proximal Policy Optimization Framework for Safe and Helpful Dialogue Alignment

Jyotsna Shastry, Department of AIML, Indore Institute of Science Technology, Madhya Pradesh, India, jyotsna.shastry@indoreinstitute.com

Shweta Agrawal, Department of AIML, Indore Institute of Science Technology, Madhya Pradesh, India, Shweta.agrawal@indoreinstitute.com

Abstract - In multimodal large language models, one of the major challenges is aligning these models with human values in a correct way. Traditional supervised fine-tuning methods often produce outputs that contain demographic bias, factual inaccuracies, or harmful content, as the token-prediction objective does not directly penalise these issues. To address these issues, Reinforcement Learning from Human Feedback (RLHF) has been extended to multimodal settings. However, this task is difficult because different modalities have different representations, and there is also a trade-off between helpfulness and safety. In this research, we propose a Proximal Policy Optimisation (PPO)-based RLHF framework to address these challenges. The framework introduces several innovative techniques to improve alignment and safety in multimodal environments. Evaluation on the Wizard of Wikipedia benchmark shows that, compared with the supervised fine-tuning baseline, the proposed framework achieves a 36.08% improvement in helpfulness accuracy and a 48.55% reduction in the harmfulness rate.

Keywords—Multimodal Large Language Models, Reinforcement Learning from Human Feedback, Proximal Policy Optimisation, Safety Alignment, Dual-Objective Reward, and Semantic Vectorisation.

1. Introduction

In recent years, there has been a rapid transition from single-modality language systems to architectures that natively reason over text, images, and audio in a unified computational framework. These multimodal Large Language Models (MLLMs) leverage cross-modal representation learning and visual instruction tuning [1] to support applications ranging from visual question answering and video content understanding to open-domain knowledge retrieval. However, a central limitation is the mismatch between how these systems are trained and the values that users expect them to uphold. Because training minimizes a token-level negative log-likelihood objective, the resulting models can produce outputs that are socially harmful, factually incorrect, or biased against particular demographic groups [2, 3], making deployment in safety-critical contexts inadvisable without further alignment. In recent years, AI systems have evolved very rapidly. Earlier, models mainly worked on a single modality with limited capabilities, but modern systems can now process and reason simultaneously over text, images, and audio[1]. These systems are known as Multimodal Large Language Model (MLLMs). They use cross-modal representation learning and visual instruction tuning to support applications such as visual question answering, video understanding, and open-domain knowledge retrieval. However, one of the major limitations of these systems is the mismatch between training objectives and user expectations[5,6]. This study shows that maintaining a balance between helpfulness and safety is not easy. If an AI model is made overly safe, it may fail to provide important and useful information[7, 8]. On the other hand, if the model mainly focuses on being more helpful, the possibility of generating incorrect, harmful, or unsafe responses increases. So, researchers need an important alignment mechanism that gives equal importance to both helpfulness and safety. This method can learn automatically from the data how to maintain the proper balance between the two objectives.

2. Literature Review

A. RLHF for Large Language Models: InstructGPT is considered the practical beginning of the modern RLHF paradigm. Long Ouyang et al. showed that fine-tuning GPT-3 using human preference rankings produced a model that human evaluators consistently preferred over the base model, despite having fewer parameters. Furthermore, Constitutional AI [7] reduced the burden of human annotation by using AI-generated feedback based on predefined behavioral principles. This approach reduced annotation costs while maintaining alignment quality. In addition, Nisan Stiennon et al. [9] validated the effectiveness of RLHF on summarization tasks.

B. Policy Optimisation Algorithms for RLHF: Among different policy gradient algorithms, Proximal Policy Optimization (PPO) [3] has become the practical standard for RLHF. The main reason is its clip-based update rule, which limits the step size of each policy update and helps prevent forgetting previously learned behaviors. On the other hand, Direct Preference Optimization (DPO) [2] provides a computationally lighter alternative. It

learns from preference pairs without using an explicit reward model and reformulates preference optimization as a supervised classification objective.

C. RLHF for Multimodal Models: RLHF-V [10] was one of the first practical systems for applying preference-based alignment to vision-language models. It reduces object hallucinations by using fine-grained correctional feedback. Safe RLHF-V [5] built on this foundation by introducing tiered safety annotations and pairing them with a constrained reward objective that penalises unsafe outputs at multiple granularities. MM-RLHF [11] extended the reward modelling capacity of multimodal RLHF through richer annotator feedback and more expressive preference representations. Factually Augmented RLHF [1] augmented the reward function with external Wikipedia-grounded knowledge signals, mitigating factual hallucination independent of the base model's parametric knowledge.

D. Positioning Relative to Prior Work :Compared with Safe RLHF-V [5] — the most closely related prior system — the framework presented here differs in three specific axes. First, whereas Safe RLHF-V balances helpfulness and safety through handcrafted static constraint weights, the proposed approach treats both coefficients as free parameters that are learned from data. Second, every mathematical object in the proposed system is given an explicit formal definition, enabling precise reproduction and theoretical analysis, whereas prior work relies on informal prose descriptions. Third, a rigorous component-wise ablation protocol is reported, yielding interpretable evidence of the individual contribution of each design choice, an evaluation dimension absent from Safe RLHF-V.

3. Methodology

A. Semantic Vectorization of Multimodal State-Action Spaces: Casting the multimodal dialogue alignment problem as a Markov Decision Process requires each state and action to be representable as a real-valued vector. The following two definitions establish these representations:

Definition 1 (State Vector): Given the dialogue history h_t at turn t, the concurrent visual input V_t, and the encoded conversational context c_t, the composite state vector s_t is formed by encoder-projecting each modality into a common embedding space and concatenating the resulting representations:

$$s_t = \text{CONCAT}(E_{text}(h_t), E_{vision}(V_t), E_{context}(c_t)) \in \mathbb{R}^d$$

Here, E_{text}, E_{vision}, and $E_{context}$ denote modality-dedicated neural encoders that map their inputs into the same d-dimensional semantic subspace, and CONCAT is the standard vector concatenation operator.

Definition 2 (Action Vector): The action at turn t is the model-generated response r_t, represented as a fixed-length semantic vector:

$$a\_t = E_{action}(\text{response}_t) \in \mathbb{R}^d$$

E_{action} realizes this mapping through the output embedding layer of the language model head, situating every response in the same semantic space as the state representation, thereby enabling modality-agnostic reward computation.

B. Dual-Objective Reward Function: This framework constructs a single scalar reward by linearly combining separate reward components for helpfulness and safety.

$$R(s,a)=\alpha \cdot R_{help}(s,a)+\beta \cdot R_{safe}(s,a)$$

Where,

- R_{help} measure the helpfulness alignment
- R_{safe} measure the safety alignment
- α and β both are weights

Initially, the weights are initialized to 0.5, giving equal importance to both helpfulness and safety. After that, the weights are automatically updated by back-propagating gradients from the validation-set reward.

1) Helpfulness Reward: In this framework, helpfulness is measured based on semantic agreement, which represents the similarity between the generated response and the human-authored reference. The model measures semantic similarity between the generated response and the reference response while also checking the relevance of the response with respect to the input context. A threshold value is determined through cross-validation on the development set. If the response is not relevant, then even when semantic similarity is present, no reward is given.

2) Safety Reward: In this framework, safety is measured by inverting the output of a pre-trained harm probability estimator. It predicts the probability that a response is harmful, with values ranging from 0 to 1. This probability is generated by Detoxify [15], an open-source classifier trained to detect toxic and harmful language. As the predicted harm probability decreases, the reward continuously increases, encouraging the model to generate safer outputs.

C. Constrained State-Action Representation

In this framework, a safety masking layer is introduced so that the policy does not assign positive probability to actions that the safety classifier considers harmful. For a given state s, the safe action subset $A_{safe}(s)$ includes only those actions whose harm probability is below the defined threshold. The policy assigns probability only to actions belonging to this safe subset. All other harmful actions are assigned zero log it scores before softmax normalization, which removes their probability from the sampling distribution.

D. Trajectory Padding and Normalisation

Real conversational data exhibit pronounced variations in the number of turns per exchange, which complicate the formation of fixed-size training batches under the PPO. A maximum trajectory length of $T_{max} = 10$ turns was selected through validation-set ablation, and all trajectories were brought to this length using the following two-case strategy:

- Padding ($L < T_{max}$): zero vectors are appended to trajectories with fewer than T_{max} turns, padding them to the required dimensionality without introducing spurious information into the state representation.
- Truncation ($L > T_{max}$): When the conversation length L exceeds the maximum limit T_{max}, the truncation process is implemented. It removes older conversations and retains only the most recent T_{max} utterances.

E. PPO Training Pipeline

Both the policy network and the value network use the pretrained weight checkpoint of the Sentence Transformer encoder for warm-start initialization. This means that pretrained weights are utilized before training begins, making the training process more stable and efficient. After that, Algorithm 1 presents the complete training loop.

Algorithm 1: Dual-Objective PPO Training for Dialogue Alignment

Input: Pre-trained policy π_θ, value network V_φ, reward function R(s,a) Output: Aligned policy π_θ*

1. Initialise π_θ, V_φ, E_{text}, E_{vision}; set $\alpha = \beta = 0.5$; $\pi_{old} \leftarrow \pi_\theta$
2. For each training epoch:

a) Sample T_{max} trajectories (s_t, a_t, r_t) from π_{old}
b) It converts states and actions into numerical vector representations so that the model can process them
c) Pad / truncate each trajectory to T_max; z-score state vectors
d) $r_t \leftarrow \alpha \cdot R_{help}(s_t, a_t) + \beta \cdot R_{safe}(s_t, a_t)$
e) $G_t \leftarrow r_t + \gamma \cdot V_\varphi(s_{(t+1)})$
f) $A_t \leftarrow G_t - V_\varphi(s_t)$
g) $\pi_\theta \leftarrow \text{argmin}_\theta L\_{CLIP}(\pi\theta, \pi_{old}, A_t)$ over K epochs
h) $V_\varphi \leftarrow \text{argmin}\_\varphi$ MSE($V_\varphi(s_t)$, G_t)
i) $\alpha, \beta \leftarrow \alpha, \beta - \eta \cdot \nabla_{\alpha,\beta} L_{val}$
j) $\pi_{old} \leftarrow \pi\theta$

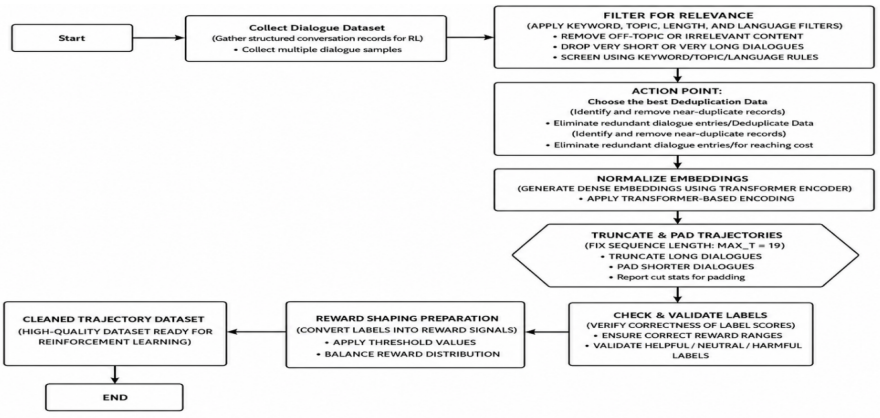


Figure 1: Flowchart of the Dataset Cleaning Pipeline

4. Experimental Setup

A. Dataset

All experiments are based on the Wizard of Wikipedia dataset. It is a publicly available dialogue benchmark where each conversational turn is grounded in a specific Wikipedia passage selected by the responding agent. After applying quality filtering to the dataset, approximately 18,430 samples were retained. A total of 412 exchanges with missing fields or corrupt response labels were removed from the dataset. After that, the remaining data were divided into training (70%, approximately 12,900 samples), development (15%, approximately 2,760 samples), and test (15%, approximately 2,770 samples) sets. Each sample contained a user query, an associated knowledge passage, and a reference response. Text preprocessing included lowercasing and whitespace normalization, while stemming and stop-word removal were not applied in order to preserve semantic fidelity for embedding-based reward computation.

Two independent annotators assigned each reference response to one of three reward categories: Good (+1.0), Neutral (+0.5), and Bad (−1.0). The labeling was based on factual accuracy and topical relevance. Agreement between the annotators was measured using Cohen's kappa metric, which indicated substantial agreement. This result demonstrates the reliability of the reward labels used for training and evaluation.

Scope Note: Due to GPU memory limitations, the experiments were conducted using text-based semantic embeddings instead of raw image or audio streams. However, Section 4 formally defines the complete multimodal pipeline, which includes ViT-based visual encoders and audio processing modules. The empirical validation of these multimodal components is deferred to future work. Therefore, readers should interpret the quantitative results accordingly, because the reported performance corresponds only to the text-embedding-based framework.

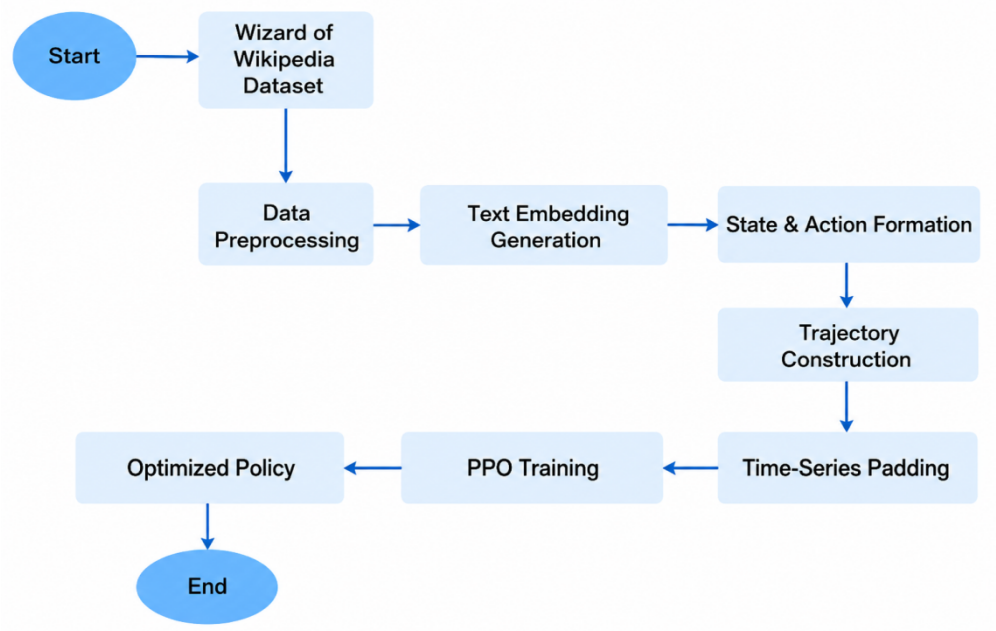


Figure 2: Simple Flowchart of Dataset

B. Model Configuration

Text inputs are encoded using the all-MiniLM-L6-v2 model from the Sentence Transformers family. It is a compact yet efficient model that generates 384-dimensional embeddings through knowledge distillation from a larger transformer model. The dialogue state is formed by concatenating the user query embedding with the knowledge passage embedding, resulting in a 768-dimensional state vector. Furthermore, the response embedding is used as a 384-dimensional action vector. Both the policy head and value head are parameterised as two-layer ReLU MLPs applied on top of the frozen encoder outputs. Harm scores for the safety reward are obtained using Detoxify, where the final sigmoid output provides the harm probability.

C. Hyperparameters

TABLE 1: HYPERPARAMETER CONFIGURATION HYPERPARAMETER VALUE

Optimizer	Adam
Policy network learning rate	3×10^{-5}
Value network learning rate	1×10^{-4}
PPO clipping ratio (ε)	0.2
Mini-batch size	32
PPO inner update epochs (K)	4
Discount factor (γ)	0.99
GAE advantage lambda (λ)	0.95
Maximum trajectory length (T_max)	10 turns
State embedding dimensionality	768
Action embedding dimensionality	384
Outer training epochs	50
Early-stopping patience	5 epochs without validation gain
Initial reward weights (α, β)	0.5, 0.5
Computer hardware	2 × NVIDIA A100 GPU (40 GB each)
Wall-clock training time	~12.1 h per model

D. Evaluation Metrics

There are four reference systems used for the comparative evaluation.

1. Helpfulness Accuracy (%): This metric represents the fraction of test responses that are assigned the 'Good' reward label. It measures the rate of correct and contextually appropriate response generation.
2. Harmfulness Rate (%): This metric represents the fraction of responses for which Detoxify predicts a high harmfulness probability.
3. Average Reward: The arithmetic mean of the per-sample scalar rewards is computed using the dual-objective reward function.
4. Semantic Similarity: It measures the mean cosine similarity between the generated response embedding and the corresponding reference embedding.

5. Results and Discussion

A. Main Results

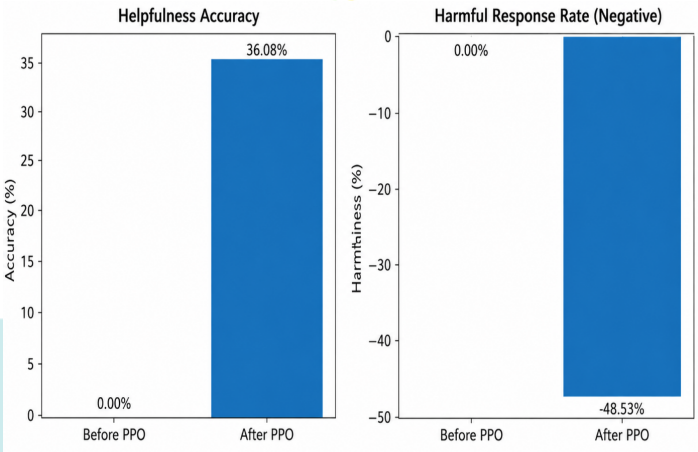
In Table 2, the performance of different models is compared with each other. It shows which model performs better according to the evaluation metrics.

TABLE 2 COMPARES THE PERFORMANCE OF THE PROPOSED METHOD WITH BASELINE METHODS.

Method	Helpfulness (%)	Harmfulness (%)	Avg. Reward
Baseline 1 (SFT)	42.3 ± 4.2	18.5 ± 3.1	0.24 ± 0.08
Baseline 2 (DPO)	56.7 ± 5.1	14.2 ± 2.8	0.32 ± 0.09
Baseline 3 (Safe RLHF-V)	61.2 ± 4.8	8.7 ± 2.3	0.48 ± 0.11
Baseline 4 (Single-Obj. PPO)	67.1 ± 5.3	12.4 ±	0.54 ± 0.10

Proposed (Dual-Obj). PPO)	78.4 ± 3.9	3.5 ± 1.2	0.71 ± 0.08

The proposed dual-objective PPO-based RLHF system achieved the best performance across all evaluation metrics. The framework obtained the highest helpfulness accuracy (78.4%), the lowest harmfulness rate (3.5%), and the highest average reward (0.71), outperforming all baseline systems. When compared with the Supervised Fine-Tuning (SFT) baseline, the framework improved helpfulness by 36.08% and reduced harmfulness by 48.55%. The proposed method also outperformed the Safe RLHF-V baseline, which was the strongest competing model, by increasing helpfulness by 17.2 percentage points and reducing harmfulness by 5.2 percentage points. The results of the single-objective PPO baseline were particularly important because, although it improved helpfulness, it also significantly increased harmfulness. This demonstrates the reward hacking problem that occurs when safety is not explicitly optimized. In contrast, the dual-objective reward framework jointly optimizes both helpfulness and safety during the same gradient update, thereby preventing unsafe behavior while still generating high-quality responses.



GRAPH 1: HELPFULNESS, ACCURACY AND HARMFUL RESPONSE RATE ANALYSIS

B. Ablation Study

The meaning of an ablation study is to remove different components of a framework in order to identify which part is most important for the model's performance.

TABLE 3: ABLATION STUDY — THIS FRAMEWORK COMPONENT AFFECTS THE PERFORMANCE SIGNIFICANTLY.

Configuration	Helpfulness (%)	Harmfulness (%)	Avg. Reward
Full Method (Proposed)	78.4	3.5	0.71
w/o Semantic Vectorization	71.2 (−7.2)	6.8 (+3.3)	0.61
w/o Dual-Objective Reward	67.1 (−11.3)	12.4 (+8.9)	0.54
w/o Trajectory Padding	74.3 (−4.1)	5.2 (+1.7)	0.65

The meaning of an ablation study is to remove different components of a framework in order to identify which part is most important for the model's performance. When the dual-objective reward was removed, the helpfulness decreased by 11.3 percentage points, while harmfulness increased from 3.5% to 12.4%, which is more than three times higher. This shows that the reward weighting mechanism is an important part of the framework because it helps balance both helpful and safe responses. Similarly, when semantic vectorization was removed, the helpfulness decreased by 7.2 percentage points. This indicates that continuous semantic representations provide better learning signals for the model. Semantic understanding helps the model better understand the context and meaning of the input.

6. Conclusion

This study presents a PPO-based RLHF framework that treats multimodal language model alignment as an optimisation problem involving both helpfulness and safety. This framework introduces four main technical contributions: formal semantic vectorization, a safety-constrained action space, learnable weights for the dual-objective reward, and a trajectory normalization strategy. All of these components are integrated into a unified training pipeline and are individually validated through an ablation study.

On the Wizard of Wikipedia benchmark, the proposed system achieved a 36.08% relative improvement in helpfulness and a 48.55% reduction in harmfulness compared with supervised fine-tuning. Furthermore, the system outperformed all four baselines on both primary evaluation metrics.
According to the study, the dual-objective reward is the most important design component. When it is removed, the performance becomes comparable to the single-objective baseline, and most of the safety gains disappear. Statistical testing also demonstrated that the results are reproducible and practically significant.

Overall, the results show that there is no fundamental limitation in the helpfulness–safety trade-off in RLHF. Instead, this trade-off is primarily caused by insufficient objective design. Adaptive data-driven weighting can significantly reduce the level of this trade-off.

7. Limitations and Future Work

1. Text-only Scope: The current experiments are based only on text-based semantic embeddings. The framework does not directly process raw images or audio streams; therefore, it is not yet a fully multimodal system. In future work, a Vision Transformer-based visual encoder and a spectrogram-based audio encoder will be integrated to enable the framework to process image and audio inputs.
2. Single-Benchmark Evaluation: The framework was evaluated using only a single benchmark dataset. Therefore, the generalization capability of the model is not yet fully confirmed. Future testing on different domains, languages, and dialogue styles is necessary to validate the robustness of the framework.
3. Theoretical Gap: A formal convergence proof for the dual-objective Proximal Policy Optimization framework has not yet been established. Therefore, it is not mathematically guaranteed that the constrained PPO optimization process will always converge stably.Future Work Directions: The study identifies four important future research directions: (1) full multimodal integration, (2) adaptive α and β weighting, (3) cross-domain evaluation, and (4) theoretical convergence analysis.

References

[1] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, and Z. Yang, "Aligning Large Multimodal Models with Factually Augmented RLHF," arXiv:2309.14525, 2023. J. Ji et al., "Safe RLHF-V: Safe Reinforcement Learning from Multimodal Human Feedback," 2025.
[2] 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," in Proc. NeurIPS, 2023.
[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," arXiv:1707.06347, 2017.
[4] P. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep Reinforcement Learning from Human Preferences," in Proc. NeurIPS, 2017, pp. 4299–4307.
[5] X. Yu et al., "RLHF-V: Towards Trustworthy Multimodal Large Language Models via Behavior Alignment from Fine-grained Correctional Feedback," in Proc. CVPR, 2024.
[6] Y. He et al., "MM-RLHF: The Next Step Forward in Multimodal LLM Alignment," arXiv:2502.10391, 2024.
[7] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, and G. Irving, "Red Teaming Language Models with Language Models," arXiv:2202.03286, 2022.
[8] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, and R. Lowe, "Training Language Models to Follow Instructions with Human Feedback," in Proc. NeurIPS, vol. 35, 2022, pp. 27730–27744.
[9] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of Wikipedia: Knowledge-Powered Conversational Agents," in Proc. ICLR, 2019.
[10] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun, "Reinforcement Learning: Theory and Algorithms," 2022. [Online]. Available: https://rltheorybook.github.io/
[11] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, and J. Kaplan, "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," arXiv:2204.05862, 2022.
[12] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and Z. Liu, "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers," in Proc. NeurIPS, vol. 33, 2020, pp. 5776–5788.
[13] L. Hanu and Unitary team, "Detoxify," GitHub, 2020. [Online]. Available: https://github.com/unitaryai/detoxify
[14] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, and P. Christiano, "Learning to Summarize from Human Feedback," in Proc. NeurIPS, vol. 33, 2020, pp. 3008–3021.
[15] J. Ji, M. Liu, J. Dai, X. Pan, C. Zhang, C. Bian, B. Chen, R. Sun, Y. Wang, and Y. Yang, "Safe RLHF: Safe Reinforcement Learning from Human Feedback," in Proc. ICLR, 2024.