RPO: Retrieval Preference Optimization for Robust Retrieval-Augmented Generation

Shi-Qi Yan[1], **Quan Liu**[2] and **Zhen-Hua Ling**[1]
[1]National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China
[2]State Key Laboratory of Cognitive Intelligence, iFLYTEK Research
sqyan01@mail.ustc.edu.cn, quanliu@iflytek.com, zhling@ustc.edu.cn

Abstract

While Retrieval-Augmented Generation (RAG) has exhibited promise in utilizing external knowledge, its generation process heavily depends on the quality and accuracy of the retrieved context. Large language models (LLMs) struggle to evaluate the correctness of non-parametric knowledge retrieved externally when it differs from internal memorization, leading to *knowledge conflicts* during response generation. To this end, we introduce the **R**etrieval **P**reference **O**ptimization (RPO), a lightweight and effective alignment method to adaptively leverage multi-source knowledge based on retrieval relevance. An implicit representation of retrieval relevance is derived and incorporated into the reward model to integrate retrieval evaluation and response generation into a single model, solving the problem that previous methods necessitate the additional procedure to assess the retrieval quality. Notably, RPO is a RAG-dedicated alignment approach that quantifies the awareness of retrieval relevance in training, first overcoming mathematical obstacles. Experiments on four datasets demonstrate that RPO outperforms RAG by 4-10% in accuracy without any extra component, exhibiting its robust generalization.

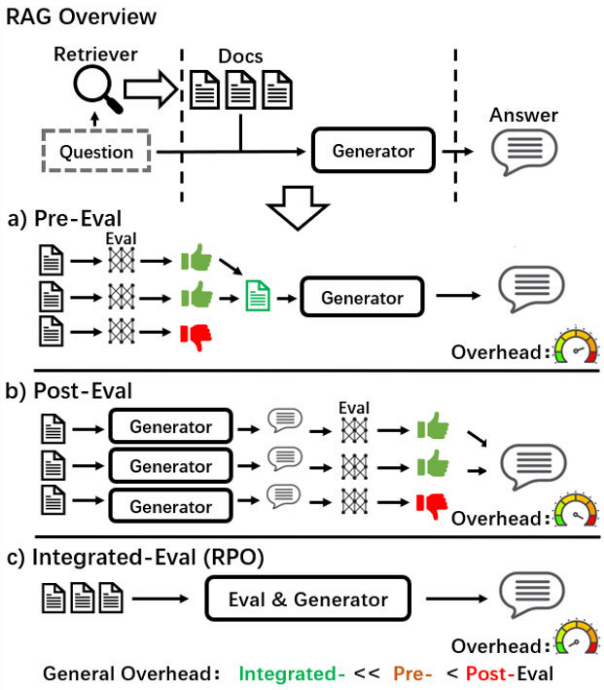

Figure 1: The figure showcases the overview of RAG and three categories of adaptive RAG, including a) Pre-Eval, b) Post-Eval, and c) Integrated-Eval approaches. The estimated computational overhead of three categories is demonstrated as well, exhibiting the efficiency of our RPO in inference.

1 Introduction

Despite the wide application in natural language processing tasks, large language models (LLMs) still struggle with knowledge-intensive tasks (Guu et al., 2020; Lewis et al., 2020a). As a general and effective approach, retrieval-augmented generation (RAG) (Lewis et al., 2020b; Izacard and Grave, 2021) involves retrieving the context related to the input query from an external corpus and integrating it for generation.

However, RAG has been found to have the potential for over-reliance on retrieval, which could unconsciously lead to hallucination, particularly when the information retrieved, also called non-parametric knowledge, conflicts with the parametric knowledge embedded within LLMs (Longpre et al., 2021; Xu et al., 2024). Specifically, RAG tends to prioritize the retrieved external context over the internal knowledge when conflicts arise (Zou et al., 2024; Xiang et al., 2024; Yan et al., 2024). Therefore, the performance of RAG depends heavily on the accuracy of the retrieval process, as inaccurate retrievals can introduce irrelevant or even harmful information, affecting the quality of generated text (Shi et al., 2023; Rony et al., 2022). To address the challenge, previous studies evaluated the quality of retrieval before (pre-eval) or after generation (post-eval). However, as

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5228–5240
July 27 - August 1, 2025

shown in figure 1, such approaches called adaptive RAG require extra processing to evaluate the value of retrieval via several API or LLM calls, leading to massive computational overhead. Meanwhile, removing part of the negative context that is assessed by the evaluator reduces the information provided for generation. It makes the generator more dependent on the evaluator, affecting the ultimate performance as well.

Considering the issues above, in this paper, we propose RPO, a **R**etrieval **P**reference **O**ptimization algorithm, aiming to enhance the robustness of LLM to multi-source knowledge by integrating retrieval evaluation in generation through reinforcement learning. A comprehensive theoretical analysis is first conducted to highlight the technical limitations of previous preference optimization algorithms (Ouyang et al., 2022; Rafailov et al., 2023; Zhang et al., 2024) in the context of the RAG scenario. We mathematically prove the limitations of the previous methods, which violate the objective of adaptive RAG, which is to select the correct answer both before and after retrieval. When conflict is involved between parametric and non-parametric knowledge, an over-tendency towards the retrieved knowledge still easily arises during the generation. Building on this theory, our RPO alignment method is designed to mitigate over-reliance on retrieval by incorporating the awareness of retrieval relevance into the reward model. To strengthen the capability of conflict mitigation, RPO simulates knowledge conflict and rectifies the discernment of LLM about which type of knowledge to prioritize. First, we instructed LLM to generate answers with and without retrieval respectively, filtering the contradictory instances as knowledge conflict. In the meantime, the relevance of the retrieved context is quantified and represented implicitly. Ultimately, the calculated relevance is integrated into the reward model for alignment to adaptively reward the positive answer in the contradictory pair based on the quality of retrieval.

As shown in figure 1, RPO (Integrated-eval) integrated the evaluation of the retrieval quality with the generation, without any additional overhead, exhibiting significant efficiency. Meanwhile, results on four datasets of PopQA (Mallen et al., 2023), Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and RGB (Chen et al., 2024) show that RPO can significantly improve the performance of RAG over prior approaches, demonstrating its consistent advancements across various benchmarks.

In summary, our contributions in this paper are three-fold: 1) We propose an optimization strategy named RPO, aimed at encouraging LLMs to synchronously evaluate the retrieved context and selectively leverage non-parametric knowledge without any explicit processing during response generation. 2) We provide a mathematical proof highlighting the inadequacy of existing preference optimization strategies for direct application in RAG-based scenarios and propose a more efficient algorithm as well as a data collection method for training to address this limitation. 3) Through experimentation involving multiple LLMs and benchmarks, we validate the efficacy of our proposed RPO algorithm and showcase its consistent performance advancements.

2 Related Work

Adaptive RAG In traditional RAG (Lewis et al., 2020b) applications, the retrieved context, referred to as non-parametric knowledge, may sometimes conflict with the parametric knowledge stored in LLMs. Previous research has explored the evaluation of retrieval quality and the adaptive use of non-parametric knowledge for conflict resolution, which can be generally categorized into pre-eval and post-eval approaches. Pre-eval methods(Yoran et al., 2024; Yan et al., 2024; Wang et al., 2024) involve employing a specialized classification language model (LM) or instructing LLMs to assess retrieval quality. In contrast, post-eval methods(Asai et al., 2023; Xiang et al., 2024) entail independently generating multiple responses based on various retrieved documents and selecting the best answer as the final response. However, on the one hand, both approaches are computationally demanding and structurally complex, resulting in decreased inference efficiency. On the other hand, part of the information is removed by the evaluator, making the generator more dependent on the performance of the evaluator, which affects the ultimate performance as well.

Model Alignment In reviewing the Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) pipeline, three main phases are included: supervised fine-tuning (SFT), reward model learning, and RL optimization. After fine-tuning a pre-trained LM a pair of answers is sampled $(y_1, y_2) \sim \pi_{\text{SFT}}(y \mid x)$, crowd workers annotate the preferred one between the pair, denoted

as $y_w \succ y_l \mid x$. A latent reward model is introduced and learned afterward to quantify the preference. Ultimately, the Proximal Policy Optimization (PPO) (Schulman et al., 2017) algorithm is adopted as the objective of RL optimization. Afterward, as one of the most popular alignment strategies, DPO (Rafailov et al., 2023) involves replacing the external reward model with a closed-form expression. Instead of learning an explicit reward model, DPO reparameterizes the reward function r using a closed-form expression with the optimal policy. The computationally lightweight approach significantly eliminates the need for direct RL optimization and outperforms existing methods.

3 Task Definition

3.1 RAG Formulation

To answer a question x from a dataset $\mathbf{D}$ with an LLM π, RAG requires the retrieved context R as the supplementary material before response generation. In most situations, the first stage of the system is to retrieve multiple relevant documents $D^r = \{D_1^r, ...D_K^r\}$ from an accessible corpus $\mathbb{C}$, which then serve as supplementary input to the query for the LLM generation. Thus the RAG task can be simplified into:

$$y_{n+p} = \pi(x, R)|_{R=D^r}, \tag{1}$$

where y_{n+p} means the answer for the question x that has access to the retrieved results, i.e., all retrieved context D^r. LLMs autonomously select either parametric or non-parametric knowledge for response generation.

3.2 Knowledge Conflict

Apart from the response that integrates retrieved information, π actually has its own potential answer with the knowledge memorized in the parameters. It can be activated by directly instructing π to generate the answer, expressed as:

$$y_p = \pi(x, R)|_{R=\phi}, \tag{2}$$

where y_p means the answer without any retrieved context, i.e. null set in the equation above, representing the response with parametric knowledge for x. Note that if the parametric knowledge and retrieved non-parametric knowledge are different, i.e., knowledge conflict arises, the generator in RAG should make a decision on which knowledge to be referred to. If the knowledge from the retrieved context is adopted, the answer would be vary from y_p. Based on this situation, we filtered the non-parametric answers y_n from y_{n+p} Ultimately, we can detect knowledge conflict and filter non-parametric answers by:

$$\mathrm{Acc}(y_n) + \mathrm{Acc}(y_p) = 1, \tag{3}$$

where $y_n \in y_{n+p}$, and the correct answer can be formulated as $\mathrm{Acc}(y) = 1$, and the incorrect one satisfies $\mathrm{Acc}(y) = 0$. Therefore, Equ. (3) indicates that only one in the pair of the answers is correct.

4 Why DPO is Limited to Apply to RAG

DPO (Rafailov et al., 2023) has shown its great performance in fine-grain optimization by aligning LLMs with the chosen ones in the preference pairs, which just meets the task requirement of the knowledge conflict. However, several concerns exist regarding the application of DPO to RAG-based tasks.

Firstly, the optimization objective of RLHF and DPO is inconsistent with the conflict-mitigating target in RAG. Considering the integrated retrieved context in the input when applied to RAG, the ultimate optimization objective of PPO-based methods such as RLHF and DPO can be formulated as :

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathbf{D}, y\sim\pi_\theta(y|x)} r(x, D^r, y) - \beta \mathbb{D}_{\mathrm{KL}}[\pi_\theta(y \mid x, D^r) \,||\, \pi_{\mathrm{ref}}(y \mid x, D^r)], \tag{4}$$

where β is the controlling hyper-parameter. π_θ and π_{ref} indicate the trainable and reference policies respectively, which are both initialized to π_{SFT}, while π_{ref} is frozen. The last term in the formulation is adopted as an extra constraint, which is significant in preventing the model from deviating to far from the original distribution. However, in the RAG application, LLMs require considerable parameter tuning to improve the distribution from the over-tendency on retrieved context. For instance, if the parametric answer is the preferred one, the ideal distribution should be aligned with $\pi_{\mathrm{ref}}(y \mid x)$, while the non-parametric answer is preferred, the target distribution should be aligned with $\pi_{\mathrm{ref}}(y \mid x, D^r)$. The constraint in the previous optimization strategies will affect the efficiency and the performance of the training methods, remaining bias on the non-parametric answers.

Secondly, the partition function within the reward model can not be canceled out. Note that DPO necessitates both positive and negative responses to have high probabilities for the same input, i.e., $(y_w, y_l) \sim \pi_{\text{SFT}}(y \mid x)$, satisfying that $\log \pi_{\text{SFT}}(y_w \mid x), \log \pi_{\text{SFT}}(y_l \mid x) > \epsilon$, where ϵ is a rather high value among the output log-probabilities of the policy. When DPO is directly applied to RAG, considering the existence of the retrieval D^r, the expression of the DPO optimizing objective can be formulated as:

$$\begin{aligned}\mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathbf{D}}\Bigg[\log \\ \sigma\Bigg(\beta \log \frac{\pi_\theta(y_w|x_w)}{\pi_{\text{ref}}(y_w|x_w)} - \beta \log \frac{\pi_\theta(y_l|x_l)}{\pi_{\text{ref}}(y_l|x_l)}\Bigg) \\ \pm \Bigg(\beta \log Z(x) - \beta \log Z(x, D^r)\Bigg)\Bigg],\end{aligned} \tag{5}$$

where $x_w = x$ and the last term is positive when the parametric answer is the positive one, while $x_w = \{x, D^r\}$ and the last term is negative when the answer with non-parametric knowledge is positive. Detailed proof can be found in Appendix A.1. Apparently, this loss function becomes complex and impractical to calculate due to the existence of the partition function.

Thirdly, over-tendency towards non-parametric knowledge is still inevitable since parametric answers are fabricated for training. Due to the issue of the partition function, the input of y_n and y_p should be the same, which does not conform to the real-world application. Prior studies have attempted to fabricate the parametric answer and pretending that it is generated with retrieved context, i.e., $(y_n, y_p) \sim \pi_{\text{SFT}}(y \mid x, D^r)$ (Zhang et al., 2024). However, the potentially significant discrepancy in likelihood between fabricated and original answers could hinder LLM convergence during training, leading to suboptimal outcomes. For instance, the situation in the inference stage widely exists where an instance satisfies $(x_{\text{inf}}, y_{p\_\text{inf}} \succ y_{n\_\text{inf}})$ but the optimized LLM still chooses the suboptimal non-parametric answer as the final response:

$$\pi_{\text{DPO}}(y_w \mid x_{\text{inf}}, D^r) < \pi_{\text{DPO}}(y_l \mid x_{\text{inf}}, D^r). \tag{6}$$

Equ. (6) suggests that despite DPO is conducted for training, the optimized policy still tends to take the dispreferred answer as the response as long as a considerable discrepancy exists between the initial preferred and dispreferred answers. Detailed proof can be found in Appendix A.2.

5 Methodology

Motivated by the challenges encountered in implementing preference optimization to RAG as illustrated above, this study aims to propose a RAG-specific approach for policy optimization. Acknowledging the discrepancy between the reinforcement learning objective of the DPO and the requirements of RAG, we first propose a new reinforcement learning objective by incorporating a representation of retrieval relevance to adaptively reward LLM based on retrieval quality. Furthermore, we outline a data collection and filtering strategy to simulate the knowledge conflict for the practical training.

5.1 Theoretically Analysis

Reward Model Since the reinforcement learning objective formulated as Equ. (4) has shown a discrepancy against the target of conflict mitigation in RAG, modifying the RL objective representation is primary and significant. In this paper, we mainly attribute the discrepancy to the absence of the retrieval rewarding. Previous studies conventionally regard retrieved context as a fixed part of the input to build the reward model, i.e., $(y_w; y_l) \mid x, R$. However, from the perspective of the entire RAG system, the retrieved context is only an intermediate variable, conditioned on the input query, which is consistent between preferred and dispreferred samples. Therefore, we suppose that the reward model in RAG should reward not only a preferred answer, but also a preferred retrieval, i.e., $(y_w, R_w; y_l, R_l) \mid x$. Ultimately, the RL objective can be formulated as:

$$\begin{aligned}\max_{\pi_\theta} &\mathbb{E}_{x\sim\mathbf{D}, y\sim\pi_\theta(y|x,R)} r(x, y, R) \\ &- \beta \mathbb{D}_{\text{KL}}[\pi_\theta(y, R \mid x) \,||\, \pi_{\text{ref}}(y, R \mid x)].\end{aligned} \tag{7}$$

Similar to the derivation of the reward model in the DPO strategy, we can get the reward model formulation in our RPO:

$$\begin{aligned}r(x, y, R) =& \beta \log \frac{\pi(y \mid x, R)}{\pi_{\text{ref}}(y \mid x, R)} \\ &+ \beta \log \frac{\pi(R \mid x)}{\pi_{\text{ref}}(R \mid x)} + \beta \log Y(x),\end{aligned} \tag{8}$$

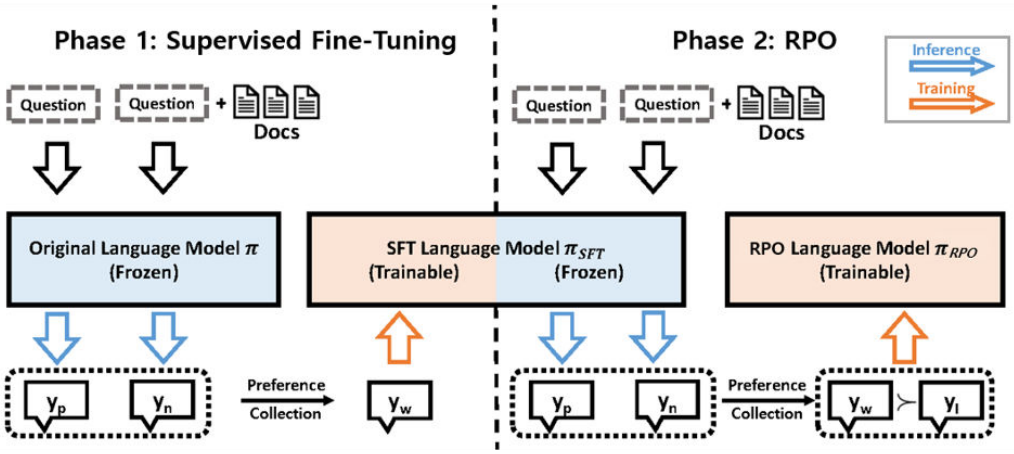

Figure 2: An overview of RPO at training. In phase 1, given a question and the retrieved documents, two answers (y_p, y_n) are generated by the frozen language model π. After comparing with the golden answers, instances that involve knowledge conflict are filtered for supervised fine-tuning. In phase two, the fine-tuned LLM is prompted to generate a pair of answers again, and the instances with knowledge conflict are filtered as the training set of RPO.

where $Y(x)$ is the partition function, the details about the reward model can be found in Appendix A.3.

Length Normalization Previous studies have observed the tendency of LLMs to be influenced by the length bias during DPO. In RPO, since retrieved context is generally much longer than the response, the length of the retrieved context could greatly affect the reward model, raising the length bias To mitigate the excessive impact of the retrieval-awareness term, and overcome the length bias of LLMs, we utilized the average log probabilities as a part of the reward. Substituting the length normalization in the reward model representation, the ultimate RPO training objective can be written as:

$$\mathcal{L}_{\text{RPO}} = -\mathbb{E}\Bigg[\log\sigma\Bigg(\underbrace{\beta\log\frac{\pi_\theta(y_w|x,D^r)}{\pi_{\text{ref}}(y_w|x,D^r)}}_{\text{(a)preferred generation reward}} \underbrace{-\beta\log\frac{\pi_\theta(y_l|x,D^r)}{\pi_{\text{ref}}(y_l|x,D^r)}}_{\text{(b)dispreferred generation reward}} \underbrace{\pm\frac{\beta}{|D^r|}\log\frac{\pi_\theta(D^r\mid x)}{\pi_{\text{ref}}(D^r\mid x)}}_{\text{(c)retreival reward}}\Bigg)\Bigg], \tag{9}$$

where the first and second terms (Equ. (9a), (9b)) represent the preferred and dispreferred reward of generation respectively, which is consistent with DPO. While the last term (Equ. (9c)) indicates the reward of the retrieved context, which is positive when the non-parametric answer y_n is preferred against the parametric answer y_p, i.e., $y_n \succ y_p$, and negative when the parametric answer is preferred, i.e., $y_p \succ y_n$.

5.2 Training Overview

In this section, we illustrate how to collect, filter, and formulate data for SFT and preference optimization. Figure 2 and Algorithm 1 present an overview of RPO at training. Each example is comprised of a query and a corresponding Wikipedia page that can answer the question and has one or more short spans from the annotated passage containing the actual answer.

Preference Pairs Collection We first construct the preference pairs adopted for supervised fine-tuning (SFT) and RPO, aimed at enhancing the model's awareness to leverage retrieved non-parametric knowledge adaptively. Given an instance from the dataset $(x, y) \in \mathbf{D}$, we respectively instruct the model to generate responses with and without retrieval (y_{n+p}, y_p) as illustrated in Section 3.2. Two subsets sampled from $\mathbf{D}$ are constructed to collect preference pairs. In the first subset $\mathbf{D}^1$, our goal is to continually enhance the model's ability to read and comprehend the retrieved context. Instances are sampled where the model fails to answer the questions directly, while correctly generating the responses with retrieval, i.e., $\text{Acc}(y_{n+p}) > \text{Acc}(y_p)$. To further confirm that y_{n+p} refers to the retrieved knowledge, i.e. $y_{n+p} = y_n$, we solely select samples where the ground truths are contained in the retrieved context. The second subset $\mathbf{D}^2$ focuses on mitigating the over-reliance of the model on the retrieved knowledge. We select the instances where the model could have responded correctly while being affected by the retrieved knowledge and generating incorrect an-

Algorithm 1: RPO Training Procedure

Model : π
Dataset(D) : $\mathcal{X}$ (Input Questions), $\mathcal{Y}$ (Output Labels), $\mathbb{C} = \{D_1, D_2, ..., D_N\}$ (Documents)
Output : π_{RPO} (Optimized Policy)

// Supervised Fine-Tuning
foreach $(x, y) \in (\mathcal{X}, \mathcal{Y})$ **do**
    $y_p = \pi(x)$
    $y_{n+p} = \pi(x, D^r)$, $D^r = \{D_j^r, j = 1, 2, ...K\}$ = Retriever(x)
end
$\mathbf{D}_{\text{SFT}}$ =Conflict_Collection($\mathbf{D}$, Condition:Acc(y_{n+p}) + Acc(y_p) = 1)
π_{SFT} =Supervised_FineTuning($\pi, \mathbf{D}_{\text{SFT}}$)
// Retrieval Preference Optimization
foreach $(x, y) \in (\mathcal{X}, \mathcal{Y})$ **do**
    $y_p = \pi_{\text{SFT}}(x)$
    $y_{n+p} = \pi_{\text{SFT}}(x, D^r)$, $D^r = \{D_j^r, j = 1, 2, ...K\}$ = Retriever(x)
end
$\mathbf{D}_{\text{RPO}}$ =Conflict_Collection($\mathbf{D}$, Condition:Acc(y_{n+p}) + Acc(y_p) = 1)
π_{RPO} =RPO($\pi_{\text{SFT}}, \mathbf{D}_{\text{RPO}}$)

swers, i.e., Acc(y_p) > Acc(y_{n+p}). Note that interference due to incorrectness is caused by the introduced non-parametric knowledge, y_{n+p} can be approximately regarded as a non-parametric answer y_n. It helps the model to reconsider whether to utilize the non-parametric knowledge before generation. Ultimately, combine both subsets and obtain the training set, $\mathbf{D}_{train} = \mathbf{D}^1 \cup \mathbf{D}^2$, which consists of samples that involve knowledge conflict.

Supervised Fine-Tuning In this stage, we perform SFT utilizing the instances that are collected with the methods in Section 5.2, obtaining the subset $\mathbf{D}_{\text{SFT}}$. Despite preference pairs are not required in the SFT stage, the subset is constructed only to collect knowledge conflict. Since only one between parametric and non-parametric sources of the instances in $\mathbf{D}_{\text{SFT}}$ contains the correct knowledge, the model must determine which knowledge to rely on. Therefore, SFT helps the model to preliminarily raise awareness of evaluating the quality of retrieval to support its decision.

Retrieval Preference Optimization As the previous illustration reveals, LLMs generally exhibit confusion and hallucination when accessing a context that contains different information than parametric knowledge. To address this issue, we propose the Retrieval Preference Optimization (RPO) training strategy, enhancing the awareness of LLMs to focus on the retrieved context during response generation. In detail, similar data filtering processing illustrated in Section 5.2 is adopted to the dataset again with the fine-tuned policy π_{SFT}. Meanwhile, which of the answers within the (y_p, y_{n+p}) pairs will be preferred is annotated by their accuracy. The selected dataset through the SFT policy utilized for subsequent training is denoted as $\mathbf{D}_{\text{RPO}}$ Eventually, we conduct the RPO strategy by reducing the loss demonstrated in Equ. (9). In this approach, we obtain the ultimate policy denoted as π_{RPO}, which implicitly conducts an integrated evaluation on retrieval within the generation.

6 Experiments

We conducted experiments to extensively demonstrate RPO's advancement and adaptability to RAG-based approaches and their generalizability across various tasks.

6.1 Tasks, Datasets and Metrics

RPO was evaluated on four datasets, including **PopQA** (Mallen et al., 2023), **NQ** (Kwiatkowski et al., 2019), **RGB** (Chen et al., 2024), and **TriviaQA** (Joshi et al., 2017). Following previous work, accuracy was adopted as the evaluation metric for the benchmarks. On the one hand, the same metrics are used because our proposed method is comparable to previous studies. On the other hand, the accuracy metric objectively measures the accuracy of the knowledge within generated responses, which appropriately represents the performance of

Table 1: Overall evaluation results on the test sets of four datasets. Results are separated based on the generation LLMs. The Column Adaptive Category indicates the category of the method if it belongs to adaptive RAG. # API/LM Calls means the number of times that an API or an LM is called during an inference. **Bold** numbers indicate the best performance among all methods and LLMs. †indicates that due to the cost, only a part of the test set is evaluated. * indicates the results that are directly cited from the papers, otherwise results are reproduced by us with the consistent retrieval results.

Method	Adaptive Category	#API/LM calls	PopQA (Accuracy)	NQ (Accuracy)	TriviaQA (Accuracy)	RGB (Accuracy)
			Others			
$\text{RAG}_{\text{ChatGPT}}$†	-	1	50.8	41.8	65.7	99.3
AstuteRAG	Pre-Eval	2-4	42.1	51.5	47.6	94.6
			LLaMA2-7B			
RAG	-	1	48.8	22.0	52.5	91.6
RAG + SFT	-	1	51.3	36.0	54.3	94.6
RAG + DPO	-	1	53.6	43.5	51.7	96.3
CRAG†	Pre-Eval	6	54.9	38.4	59.6	92.0
Self-RAG	Post-Eval	2-11	54.9	42.4	**68.9**	92.6
RPO	Integrated-Eval	1	**55.8**	**45.3**	57.6	**97.3**
			LLaMA3-8B-instruct			
RAG	-	1	59.0	41.3	65.8	96.3
InstructRAG	-	1	65.0	46.7	65.1	99.3
Self-RAG*	Post-Eval	2-11	55.8	42.8	71.4	-
RPO	Integrated-Eval	1	**65.4**	**51.9**	**74.4**	**100.0**

methods in knowledge-intensive tasks.

6.2 Baselines

We primarily compared RPO with previous RAG-based baselines, which can be divided into three categories according to the base model, including:

LLaMA2-7B approaches utilized the vanilla or instruction-tuned LLaMA2-7B model for response generation. (1) RAG + SFT directly tuned the model with the instances that involve knowledge conflict. (2) RAG + DPO tuned the model with SFT in phase 1, while tuning the model with DPO rather than RPO in Phase 2. Conflict collection is implemented in both SFT and DPO before training to ensure comparability. (3) Self-RAG (Asai et al., 2023) that tuned the LLaMA2 on the instruction-tuning data containing several sets of reflection tokens which were labeled by GPT-4 (OpenAI, 2023), while (4) CRAG (Yan et al., 2024) that evaluated the quality of the retrieval and selectively corrected the retrieved context with the web search.

LLaMA3-8B-Instruct approaches generated the response with LLaMA3-8B-Instruct. (1) InstructRAG (Wei et al., 2024) proposes a instruction-tuning method, while (2) Self-RAG are along with the methods above except the base model. Notably, results on Self-RAG with * indicate that the results are directly cited from the previous paper.

Commercial APIs refers to the approaches that import commercial LLMs for text generation. We introduce the methods driven by commercial APIs for reference to benchmark the broader effectiveness and efficiency of our proposed RPO. Specifically, AstuteRAG (Wang et al., 2024) was reproduced in this experiment on ChatGPT, which iteratively filtered and revised the knowledge before generation.

6.3 Results

Table 1 presents the results on four datasets. We briefly mark the categories of the listed adaptive RAG methods in the table. To showcase the efficiency of RPO in computational overhead during the inference phase, the estimated API call or LLM inference times are presented as well. From these results, we can conclude the following findings:

First, the proposed method significantly outperformed previous baselines that involve adaptive retrieval, reaching state-of-the-art. Specifically, as shown in table 1, RPO outperformed RAG by

Table 2: Ablation study for removing retrieval-awareness, preference optimization, and SFT phases respectively on the PopQA dataset in terms of accuracy. w/o RR means that the retrieval reward term is removed for optimization, while w/o PO means that the model is trained without preference optimization.

	PopQA	NQ	TriviaQA	RGB
LLaMA2-7B-hf				
RPO	55.8	45.3	57.6	97.3
RPO w/o RR	53.6	43.5	51.7	96.3
RPO w/o PO	51.3	36.0	54.3	94.6
RPO w/o SFT	52.5	34.9	50.1	90.6

margins of 6.4% accuracy on PopQA, 10.6% accuracy on NQ, 8.6% accuracy on TriviaQA, and 3.7% accuracy on RGB when based on *LLaMA3-8B-instruct*, as well as by margins of 7.0% accuracy on PopQA, 23.3% accuracy on NQ, 5.1% accuracy on TriviaQA, and 5.7% on RGB when based on *LLaMA2-hf-7b*. Compared with the currently advanced adaptive RAG methods, RPO has generally outperformed in all the benchmarks. The advancements in our method greatly illustrate the effectiveness of preference optimization, showing the significance of overcoming the knowledge conflict.

Second, the proposed method exhibited greater computational efficiency, providing a practical solution in the real-world application for knowledge conflict mitigating. It can be seen that either preeval or post-eval approaches require multiple calls of API or LMs within a single inference. Compared to the previous adaptive RAG, the retrieval evaluation is performed synchronously through generation. Meanwhile, even better results are obtained, further illustrating the efficacy of our RPO.

6.4 Ablation Study

Given that our training pipeline incorporates two distinct phases—supervised fine-tuning and preference optimization and both of which contribute to enhancing retrieval awareness and mitigating knowledge conflict, we conduct ablation studies to evaluate the individual contribution of each phase within our RPO framework. The fine-tuning and preference optimization phases are removed specifically in the experiment and the results are evaluated on the benchmarks. It is worth noting that, since the retrieval reward term in Equ. (9) is the biggest difference between DPO and RPO, RPO without the retrieval reward term (RR) can be equivalent to a DPO model. Similarly, RPO without preference optimization represents models trained solely via supervised fine-tuning, omitting the subsequent alignment stage. Results in Table 2 demonstrate that the performance dropped when removing either phases, revealing the significance.

Table 3: The robustness of each training strategy to low-quality retrieval in the PopQA dataset, where all retrieval information is incorrect.

	Acc in Low-Quality Retrieval
LLaMA2-7B-hf	
RAG	18.6 (0.0%)
SFT	19.5 (+4.8%)
DPO	19.3 (+3.7%)
RPO	**23.5 (+26.3%)**

Table 4: Comparison results between RPO with and without data filtering during SFT phase.

	PopQA	NQ	TriviaQA	RGB
LLaMA2-7B-hf	48.8	22.0	52.5	91.6
π_{SFT} w/o filtering	46.9	38.2	48.8	80.0
π_{SFT} with filtering	51.3	36.0	54.3	94.6

6.5 Robustness to Low-Quality Retrieval

As illustrated above, one of the primary objectives in this paper is to improve the ability of LLMs to select accurate information amidst knowledge conflicts. It frequently occurs in a low-quality retrieval environment, posing significant challenges for prior methods. Therefore, to further evaluate the robustness of RPO to low-quality retrieval, we simulate this environment by assessing the performance of LLMs when provided *only* with incorrect information. Results in Table 3 reveal the performance degradation of various methods under the condition of erroneous retrieval context in PopQA. Although all methods inevitably suffer from performance degradation, our RPO still maintains a superior performance. The experiments further demonstrate that unlike DPO, which exhibits limitations and potential biases when applied to RAG, RPO can effectively evaluate the correctness of the retrieved context during the response generation.

6.6 Impact of Training Set Filtering

In phase 1 of the training stage, supervised fine-tuning is introduced for the preliminary training. Notably, the training set is filtered, only the in-

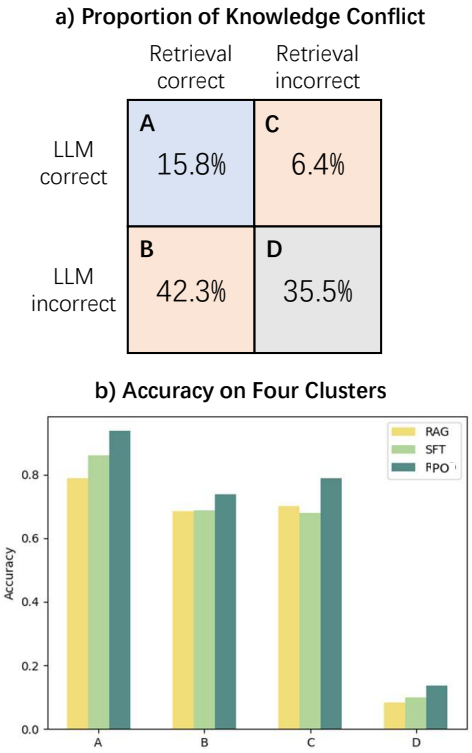

Figure 3: Proportion of four clusters in PopQA and the corresponding accuracy scores on LLaMA2-7B.

stances that involve knowledge conflict are selected for supervised fine-tuning. We hypothesize that LLMs possess the inherent ability to assess retrieval quality while generating responses, albeit not activated yet. Therefore, the operation is solely intended to enhance the retrieval awareness of LLMs, rather than to learn more knowledge. In fact, the experimental results in table 4 reveal that the fine-tuned LLM without data filtering significantly underperformed, even worse than the original LLM before tuning, further verifying our hypothesis.

6.7 Knowledge Selection Performance

In this section, we compare RPO with previous training strategies in terms of knowledge selection performance. Further analysis is conducted on the issue of knowledge conflict before and after RPO. The results in figure 3 reveal a consistent advancement in all clusters to evaluate the knowledge and select the correct autonomously. Besides, we found that the ability of the LLM to select knowledge can be even worse after SFT. In Cluster B and C, which involve knowledge conflicts, SFT does not achieve a positive advancement, while RPO has shown a significant improvement in knowledge selection.

7 Conclusion

This paper studies the issue of knowledge conflict where parametric knowledge and retrieved non-parametric knowledge in RAG are inconsistent. Previous model alignment methods have been proved limited in the context of RAG application, leading to inadequacy and bias when knowledge conflict is involved. Therefore, a new proximal policy optimization algorithm named Retrieval Preference Optimization is proposed to adapt the RAG application. The capability of LLMs to evaluate of the retrieval is integrated into the generation with our RPO, which greatly improves the efficacy compared with previous adaptive RAG approaches. Experiments extensively demonstrate its advancement as well as generalizability across various benchmarks. Future work will continually explore a more integrated and implicit approach for retrieval evaluation to further enhance the reliability and robustness of RAG.

Limitations

While we primarily proposed to improve the RAG framework with a dedicated alignment method, whether a better reward function exists requires further study. Although we make an effort to prevent reward hacking during the experiments, the intended objective can still not be fully fulfilled. In addition, since the model is only trained on NQ, the training data could not cover various domains, leading to potential bias. Future work will further explore a more flexible and robust rewarding strategy for RAG.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *CoRR*, abs/2310.11511.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17754–17762. AAAI Press.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. 2023. Aligning language models with preferences through f-divergence minimization. In *International*

Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, pages 11546–11583. PMLR.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 874–880. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. 2022. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020a. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7052–7063. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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 *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. 2019. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *CoRR*, abs/1910.00177.

Jan Peters and Stefan Schaal. 2007. Reinforcement learning by reward-weighted regression for operational space control. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pages 745–750. ACM.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Md. Rashad Al Hasan Rony, Ricardo Usbeck, and Jens Lehmann. 2022. Dialokg: Knowledge-structure

aware task-oriented dialogue generation. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2557–2571. Association for Computational Linguistics.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arik. 2024. Astute RAG: overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. *CoRR*, abs/2410.07176.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2024. Instructrag: Instructing retrieval-augmented generation with explicit denoising. *CoRR*, abs/2406.13629.

Chong Xiang, Tong Wu, Zexuan Zhong, David A. Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust RAG against retrieval corruption. *CoRR*, abs/2405.15556.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8541–8565. Association for Computational Linguistics.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *CoRR*, abs/2401.15884.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ruizhe Zhang, Yongxin Xu, Yuzhen Xiao, Runchuan Zhu, Xinke Jiang, Xu Chu, Junfeng Zhao, and Yasha Wang. 2024. Knowpo: Knowledge-aware preference optimization for controllable knowledge selection in retrieval-augmented language models. *CoRR*, abs/2408.03297.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models. *CoRR*, abs/2402.07867.

A Detailed Poofs

A.1 Proof for Equation 5

In DPO optimization algorithm, a latent reward model $r(x, y)$ is adopted, which is consistent with RLHF. To quantify the preferences, the Bradley-Terry model is introduced, which can be written as:

$$p(y_w \succ y_l|x) = \sigma(r(x, y_w) - r(x, y_l)), \quad (10)$$

where σ is the logistic function. Therefore, given a reward model $r(y, x)$, the task can be defined as a binary classification problem and the negative log-likelihood loss can be:

$$\mathcal{L}_R(r, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim D}[\log \sigma(r(x, y_w) - r(x, y_l))]. \quad (11)$$

If DPO is directly adopted for RAG and taking y_n and y_p as the (y_w, y_l) pair, considering the influence of retrieved context D^r, the expression of reward model would get modified as:

$$r(x, y_p) = \beta \log \frac{\pi_\theta(y_p|x)}{\pi_{ref}(y_p|x)} + \beta \log Z(x); \quad (12)$$

$$r(x, y_n) = \beta \log \frac{\pi_\theta(y_n|x, D^r)}{\pi_{ref}(y_p|x, D^r)} + \beta \log Z(x, D^r). \quad (13)$$

Substituting the representation in Equ. (12) and (13) for the Bradley-Terry model in Equ. (10), it can be found that the partition function can not be canceled.

A.2 Proof for Equation 6

In order to apply DPO to RAG, fabricated answers are necessary. However, the fabricated answer is may not the candidate answers with the highest likelihood for LLMs, i.e., existing y_p, satisfying that:

$$\begin{cases} \log \pi_{\text{ref}}(y_n \mid x, D^r) > \epsilon \\ \log \pi_{\text{ref}}(y_p \mid x, D^r) < \epsilon \\ \log \pi_{\text{ref}}(y_n \mid x, D^r) - \log \pi_{\text{ref}}(y_p \mid x, D^r) > \epsilon_d, \end{cases} \quad (14)$$

where ϵ_d indicates the difference of logits between parametric output and non-parametric output, which can be massive. While π_{SFT} is π_{ref}, which is used as the reference policy in the optimization phase.

It could lead to a concern that the optimized LLMs would not converge to the optimal solution. Two aspects can theoretically interpret the conclusion. On the one hand, the proposal of the DPO reward model training strategy comes from the RL optimization objective of RLHF, as shown in Equ. (4). Therefore, due to the constraint of the KL-divergence, the distribution of the policy would not change a lot, i.e.:

$$\begin{cases} |\log \pi_\theta(y_n \mid x, D^r) - \log \pi_{\text{SFT}}(y_n \mid x, D^r)| < \epsilon_{ad} \\ |\log \pi_\theta(y_p \mid x, D^r) - \log \pi_{\text{SFT}}(y_p \mid x, D^r)| < \epsilon_{ad}, \end{cases} \quad (15)$$

where $\epsilon_{ad} > 0$ is a very limited value. Supposing a situation during the inference $(x_{\text{inf}}, y_{p\_\text{inf}} \succ y_{n\_\text{inf}})$ that can generally exist, where the parametric answer is wining, meanwhile, the distance between parametric and non-parametric is big enough so that $\epsilon_d > 2\epsilon_{ad}$, then the generator would still choose the losing one as the ultimate response:

$$\begin{aligned} \pi_\theta(y_w \mid x_{\text{inf}}, D^r) &= \pi_\theta(y_{p\_\text{inf}} \mid x, D^r) \\ &< \pi_{\text{ref}}(y_{p\_\text{inf}} \mid x_{\text{inf}}, D^r) + \epsilon_{ad} \\ &< \pi_{\text{ref}}(y_{n\_\text{inf}} \mid x_{\text{inf}}, D^r) - \epsilon_{ad} \\ &< \pi_\theta(y_{n\_\text{inf}} \mid x_{\text{inf}}, D^r) \\ &= \pi_\theta(y_l \mid x_{\text{inf}}, D^r). \end{aligned}$$

A.3 Derivation of RPO's Reward Model

Given the RL objective as Equ. (7) shows, expanding the KL-divergence Formula and derive:

$$\begin{aligned} &\max_{\pi_\theta} \mathbb{E} r(x, y, R) \\ &\quad - \beta \mathbb{D}_{\text{KL}}[\pi_\theta(y, R \mid x) \,||\, \pi_{\text{ref}}(y, R \mid x)]. \\ &= \min_{\pi_\theta} \mathbb{E}\left[\log \frac{\pi_\theta(y, R \mid x)}{\pi_{\text{ref}}(y, R \mid x)} - \frac{1}{\beta} r(x, y, R)\right], \end{aligned} \quad (16)$$

while following the previous work(Peters and Schaal, 2007; Peng et al., 2019; Korbak et al., 2022; Go et al., 2023; Rafailov et al., 2023), it is straightforward to show that the optimal solution takes the form:

$$\pi_r(y, R \mid x) = \frac{\pi_{\text{ref}}(y, R \mid x) \exp\left(\frac{1}{\beta} r(x, y, R)\right)}{Y(x)}, \quad (17)$$

where the partition function can be formulated as:

$$Y(x) = \sum_y \sum_R \pi_{\text{ref}}(y, R \mid x) \exp\left(\frac{1}{\beta} r(x, y, R)\right).$$

Based on Equ. (17), the reward model can be derived and written as:

$$r(x, y, R) = \beta \log \frac{\pi(y, R \mid x)}{\pi_{\text{ref}}(y, R \mid x)} + \beta \log Y(x). \tag{18}$$

Following the Bayes theorem, the reward model can be formulated as Equ. (8).

B Experiment Details

B.1 Details of the Datasets

RPO was evaluated on four datasets, which are in public domain and licensed for research purposes, including:

PopQA (Mallen et al., 2023) is a *short*-form generation task. Generally, only one entity of factual knowledge is expected to be answered for each single question. In our experiments, we exactly followed the setting in the previous work (Asai et al., 2023) which evaluated methods on a long-tail subset consisting of 1,399 rare entity queries whose monthly Wikipedia page views are less than 100.

Natural Questions (**NQ**) (Kwiatkowski et al., 2019) is a benchmark for question answering research that contains real user questions issued to Google search, and answers found from Wikipedia by annotators. Annotations include long answers (usually a paragraph of text) and short answers (one or more entities), which are marked as null if there is no answer on the page. Additionally, NQ contains 307,372 training examples, 7,830 examples for development, and we withold a further 7,842 examples for testing. Only short answers are adopted in our experiments.

TriviaQA (Joshi et al., 2017) is a reading comprehension dataset containing over 650K question-answer-evidence triples. TriviaQA includes 95K question-answer pairs authored by trivia enthusiasts and independently gathered evidence documents, six per question on average, that provide high quality distant supervision for answering the questions.

Retrival-Augmented Generation Benchmark (**RGB**) (Chen et al., 2024) is a benchmark that chooses to aggregate the latest news. Different basic abilities of LLMs are evaluated according to the common challenges in RAG, including noise robustness, negative rejection, information integration and counterfactual robustness.

B.2 Experimental Setup

We use the package *vllm* for inference, and the parameter settings are listed below:

```
temperature=0.0
top_p=1.0
max_tokens=100
skip_special_tokens=false
```

The model was trained on 4*A100 in our experiment, and the SFT was implemented with the hyperparameter settings below:

```
n_epochs=1
batch_size=4
gradient_accumulation_steps=32
mixed_precision=bf16
max_seq_length=2048
warmup_ratio=0.03
learning_rate=2e-5
weight_decay=0.0,
```

while RPO strictly followed the hyperparameters used in Rafailov et al. (2023).