Retrieval-DPO: Retrieval-augmented preference optimization with non-paired preference data

Anonymous ACL submission

Abstract

Aligning Large Language Models (LLMs) with human feedback is important and challenging. Rafailov et al. (2023) propose Direct Preference Optimization (DPO), a simple but effective alignment method which is reinforcement learning free. However, DPO requires paired preference data which is harder and more expensive to obtain compared to binary preference data. We propose a retrieval-based method named Retrieval-DPO to align LLMs under binary preference data situation. The core idea of our method is that learning how to align can be achieved with non-paired preference data of similar questions rather than strictly paired preference data considering the learning process of human. For instance, to teach the LLM to learn how to treat multiple perspectives, other comprehensive golden answers of similar question may have similar positive effects as the golden answer of the same question. Following this idea, we retrieve an example with opposite label from the retrieval database for a binary preference data in the training set. After the retrieval process, we get a pair of preference data but with possibly different questions and then adopt the DOVE (Bansal et al., 2024) optimization objective for the alignment. We compare Retrieval-DPO with other preference optimization algorithms which do not need paired preference data such as Kahneman-Tversky Optimization (KTO) and Unified Language Model Alignment (ULMA). We find that our method significantly outperforms KTO and ULMA on helpful-base subset of HH dataset (over 13%) and slightly outperforms KTO on harmless-base subset of HH dataset and controlled sentiment generation task. Besides, our method is not sensitive to the ratio of the number of positive examples to the number of negative examples without additional hyperparameter tuning.

1 Introduction

Alignment is considered an essential process for training large language models to be helpful, honest and harmless (Askell et al., 2021). Among the alignment methods, Reinforcement learning with Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022) has proven to be highly effective by employing reinforcement learning. However, the complexity and instability of reinforcement learning present significant challenges in practical applications. To overcome these issues, several RL-free alignment algorithms have been developed, such as Direct Preference Optimization (DPO) (Rafailov et al., 2023). DPO reparameterizes the reward function used in RLHF and optimizes the policy model without relying on reinforcement learning and a reward model. This approach simplifies the process and enhances stability.

However, alignment algorithms such as RLHF and DPO require paired preference data, which is difficult and costly to collect in real-world settings. Since 2023, researchers have pursued two main approaches to address this challenge, detailed further in section 6

- The first approach involves directly optimizing binary preferences. Notable algorithms include Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024) and Unified Language Model Alignment (ULMA) (Cai et al., 2023).

- The second approach focuses on generating paired preference data through sampling techniques, followed by applying existing paired preference optimization algorithms. Representative methods include Self-Play Fine-Tuning (SPIN) (Chen et al., 2024) and Self-Augmented Preference Optimization (SPPO) (Yin et al., 2024).

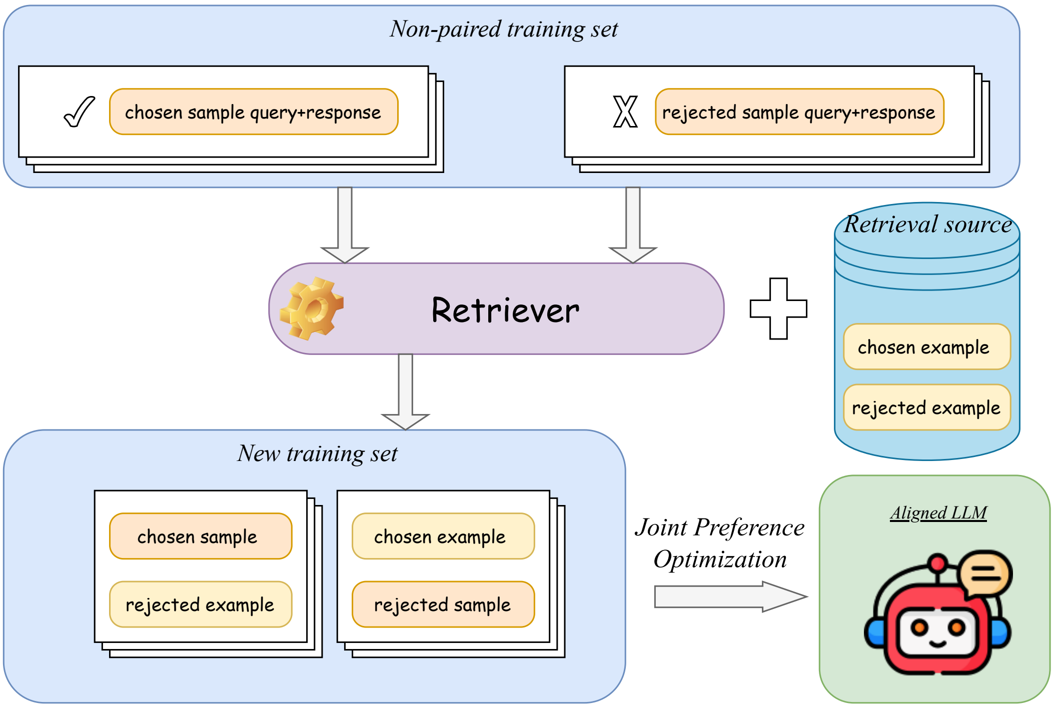

Figure 1: Overview of our Retrieval-DPO pipeline. We retrieve complementary query-response pair from the retrieval source and combine the retrieval pair with the training data to construct joint preference pairs. Finally, we perform joint preference optimization (DOVE) to get the aligned model. Q represents the query and R represents the response.

In this work, we construct preference pairs using retrieval rather than sampling. By retrieving question similar to the binary preference data from the retrieval source, we could adopt the same optimization objective as Bansal et al. (2024) to learn from joint preference. Direct application of paired preference optimization algorithms like DPO is not feasible, as we cannot ensure the presence of identical questions in the retrieval source. Our retrieval-based alignment process, which we refer to as Retrieval-DPO, involves three steps:

1. Prepare a retrieval source containing both positive and negative query-response pairs.
2. Retrieve complementary binary preference data from the retrieval source to form joint preference pairs.
3. Align LLM using joint preference optimization.

The process described above is illustrated in figure 1.

Our Retrieval-DPO offers several advantages over existing non-paired preference optimization algorithms::

1. Greater cost-effectiveness and performance: While the retrieval process requires some time and computational resources, its costs are considerably lower than those associated with the data annotation required by other methods such as DOVE. Even a simple retrieval model, without fine-tuning, can achieve satisfactory performance, especially when compared to the costs and outcomes associated with methods like ULMA and KTO. The additional retrieval costs remain manageable in the context of significant performance improvements, particularly when the retrieval source is not overly large.
2. Lower sensitivity to the ratio of the number of desirable data to the number of undesirable data. When the binary preference training set is imbalanced, if our retrieval source is of good quality we do not need to worry

about the imbalance problem. For instance, λ_D and λ_U needs to adjusted for different ratios in KTO algorithm.

Our experimental results and analysis support the advantages we mentioned above for alignment tasks including controlled sentiment generation and single-turn dialogue.

2 Background

Direct preference optimization DPO is a widely used offline paired preference optimization algorithm without the need for a reward model and reinforcement learning techniques. Under the Bradley-Terry model, DPO aims to minimize the following objective:

$$\mathcal{L}_{DPO} = -E_{(x,y_w,y_l)\sim D}(\log \sigma(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)})) \tag{1}$$

In this formula, (x, y_w, y_l) are triples from the preference dataset D, consisting of an input x, a chosen response y_w , and a rejected response y_l. The function $\pi_\theta(y|x)$ represents the model's predicted probability of response y given input x, and $\pi_{\text{ref}}(y|x)$ represents a reference probability. The hyperparameter β is a scaling factor.

Non-paired preference optimization We introduce two non-paired preference optimization which are used as baselines for our experiments here.

$$v(z; \lambda, \alpha, z_0) = \begin{cases} (z - z_0)^\alpha & \text{if } z \geq z_0 \\ -\lambda(z_0 - z)^\alpha & \text{if } z < z_0 \end{cases} \tag{2}$$

KTO starts from the Kahneman-Tversky value function (2) and replaces λ with λ_D and λ_U as hyperparameters for desirable and undesirable losses separately. Besides, KTO assumes that the reference point z_0 should be related to all possible input-output pairs rather than one data point. Combining all the above and let λ_y denote λ_D and λ_U, the KTO loss is:

$$\mathcal{L}_{\text{KTO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{x,y\sim D}[\lambda_y - v(x, y)] \tag{3}$$

where

$$\begin{aligned} r_\theta(x, y) &= \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \\ z_0 &= \mathbb{E}_{x'\sim D}\left[\text{KL}(\pi_\theta(y'|x') \| \pi_{\text{ref}}(y'|x'))\right] \\ v(x, y) &= \begin{cases} \lambda_D \sigma\left(\beta(r_\theta(x, y) - z_0)\right) & y \in \mathcal{D}^1 \\ \lambda_U \sigma\left(\beta(z_0 - r_\theta(x, y))\right) & y \in \mathcal{D}^0 \end{cases} \end{aligned} \tag{4}$$

where $\mathcal{D}^1$ and $\mathcal{D}^0$ are the chosen and rejected query-response datasets.

ULMA is inspired by point-wise DPO which is developed in the same work and use the SFT loss for the positive samples and an additional KL regularizer for the negative samples. The ULMA loss is:

$$\mathcal{L}_{\text{ULMA}}(\theta) = \sum_{(x_i,y_i,z_i)\in\mathcal{D}} -z_i \log \pi_\theta(y_i \mid x_i) - (1 - z_i) \log\left(1 - \sigma\left(r_\theta(x_i, y_i) + \beta \log Z(x_i)\right)\right) \tag{5}$$

where $r_\theta(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$

Joint preference optimization It is considerable that the preference data doesn't always exist and usually we only have supervised fine-tuning dataset $\mathcal{D} = \{(x, y)\}$. In (Bansal et al., 2024), the authors use the method of "Comparing bad apples to good oranges", which means that for two different pairs of query-response samples $(x_1, y_1), (x_2, y_2)$, we can use human or AI annotators to give the preference of this two samples, saying (x_1, y_1) is better than (x_2, y_2). Then maximizing the difference of the rewards of these two samples (although they don't have the same queries) would be reasonable. Hence we get the joint preference optimization method. We will show the details of this idea in Section 3.3.

3 Retrieval-DPO: Non-paired preference optimization with retrieval

3.1 Motivation

Compared to using paired preference data, aligning LLMs with non-paired preference data is feasible and reflects the adaptive nature of human learning. Observations of how humans learn to respond helpfully reveal that paired preference responses are not always necessary. Humans frequently encounter novel dialogues and draw on similar past experiences to formulate responses. For example, consider a scenario where a question in Table 6 asks about basketball rules. If a new question seeks the rules for baseball but only provides a rejected response, a human can adapt the previously chosen response about basketball to deliver a helpful answer by replacing basketball-specific information with baseball knowledge. Conversely, if the new question includes an chosen response, the rejected basketball response helps avoid similar unhelpful answers for baseball. Turning back to the optimization of LLMs, the discussion corresponds to joint preference optimization (Bansal et al., 2024). This adaptability in human responses mirrors the concept of joint preference optimization in LLMs, as discussed in (Bansal et al., 2024). However, if no analogous examples are available, the challenge of responding appropriately increases significantly, a difficulty also reflected in LLM optimization. Further evidence supporting this approach will be presented in Section 4.

A practical implementation of this approach involves establishing a retrieval-based system. This system would search for similar queries that have complementary binary preference data. For instance, retrieving a chosen query-response example when the binary preference signal of training data is rejected. In this work, we focus on the most straightforward method: a single retrieval followed by joint preference optimization.

3.2 Construct paired preference data with retrieval

3.2.1 Prepare the retrieval source

The initial step in our approach is to establish a retrieval source that contains binary preference data. To enhance the likelihood of retrieving relevant examples, it is advantageous to use a source with a distribution similar to that of the training data. Consequently, the training set itself is a reasonable choice and will serve as the default retrieval source in our experiments unless specified otherwise.

3.2.2 Retrieval

We utilize a retrieval pipeline analogous to the standard Retrieval-Augmented Generation (RAG) approach, as outlined in the RAG survey paper (Gao et al., 2023). This process comprises two primary steps:

1. Indexing. For simplicity, queries are encoded into vectors using a pre-trained dense encoder, specifically the Contriever model. This approach is effective given that the length of queries generally falls within the context limitations of contemporary language models. This encoding strategy will be used in all subsequent experiments.

2. Retrieval. To enhance the efficiency of our retrieval process, we use FAISS, a library for fast similarity search (Johnson et al., 2019), to locate queries that are closely related to the query.

3.3 DOVE optimization

Algorithm 1 The algorithm of Retrieval-DPO

Input: π_{ref} the reference model, $\mathcal{D}^+, \mathcal{D}^-$ the set of chosen and rejected responses
Output: π_θ the policy model

1: $\pi_\theta \leftarrow \pi_{ref}$
2: **for** prompt-response pair (x, y) in $\mathcal{D}^+ \sqcup \mathcal{D}^-$ **do**
3: Get $(\tilde{x}, \tilde{y})$ // Retrieval step
4: $\ell \leftarrow -\log\sigma(\beta_\eta \log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} - \beta_\eta \log\frac{\pi_\theta(\tilde{y}|\tilde{x})}{\pi_{ref}(\tilde{y}|\tilde{x})})$
5: $\theta \leftarrow \theta - \alpha\nabla_\theta \ell_1$ // Update parameters
6: **end for**
7: **return** π_θ

We set $\mathcal{D}^1$ and $\mathcal{D}^0$ to be the prompts with chosen responses and rejected responses. We use f to denote the retrieval map. For a given $x \in \mathcal{D}^\eta$ for $\eta \in \{0, 1\}$, we write $\tilde{x} = f(x)$ be the element we find in $\mathcal{D}^{1-\eta}$.

We follow the description of the reward function in(Rafailov et al., 2023), use $\beta \log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}$ as the reward of (x, y), as what the authors did in (Bansal et al., 2024). However, the different thing is that we don't need to annotate the preference of the pair $((x, y), (\tilde{x}, \tilde{y}))$ for $\tilde{x} = f(x)$: One is from $\mathcal{D}^1$ and the other one from $\mathcal{D}^0$, by the retrieval process x and $\tilde{x}$ should be similar with each other and we may always assume the one from $\mathcal{D}^1$ should have higher reward. Hence for a given $(x, y) \in \mathcal{D}^\eta$, and the corresponding $(\tilde{x}, \tilde{y}) \in \mathcal{D}^{1-\eta}$, we consider the loss function

$$-\log\sigma(\beta_\eta \log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} - \beta_\eta \log\frac{\pi_\theta(\tilde{y}|\tilde{x})}{\pi_{ref}(\tilde{y}|\tilde{x})})$$

where $\beta_\eta = \beta$ if $\eta = 1$ and $-\beta$ otherwise. This loss function will make the difference of the rewards for the pair $((x, y), (\tilde{x}, \tilde{y}))$ with $\tilde{x} = f(x)$ becomes larger, which is consistent with the human preference.

4 Experiments

4.1 Setup

4.1.1 Tasks, datasets, models, and training setting

We evaluate our method across two tasks: controlled sentiment generation and single-turn dialogue.

Controlled sentiment generation We follow DPO's setting and use GPT-2 as the base model. Starting from IMDB dataset (Maas et al., 2011), we employ the first eight tokens (tokenized by the GPT2-large tokenizer) of the movie review as prompts with the entire movie review serving as the target completion. We fine-tune the GPT2-large model on this dataset for a single epoch, resulting in a model capable of performing text completion tasks. For each prompt in the training set, the fine-tuned model generates four distinct completions. The sentiment of these completions is assessed using a pre-trained sentiment classifier (Hartmann et al., 2023), same as the ground-truth reward model used in the DPO paper. For the 4 generated completions, we randomly select one with positive sentiment labeled by the classifier as the chosen completion and one with negative sentiment as the rejected completion. If a prompt lacks either positive or negative completions among the four, it is excluded from the training set. Subsequently, the size of the paired perference training set is reduced to 16,056 cases from an initial count of 25,000.

Single-turn dialogue task We adopt the helpful-base and harmless-base subset of the Anthropic-HH dataset (Bai et al., 2022) and employ Llama2 (Touvron et al., 2023) as the base model.

To simulate a non-paired preference situation from the paired preference dataset, we implement a strategy similar to that used in the KTO approach. We randomly sample some examples from the training set, using only the chosen responses with their queries to form the positive examples set and the remaining examples' rejected responses with their queries to create the negative examples set. This results in a completely non-paired binary preference dataset, assuming no duplicate data exists. In practice, partially paired preference data is more common, so no further deduplication is necessary. We test three ratios (20

The instruction-following model is obtained through SFT using all chosen responses in the initial dataset for one epoch. For alignment, the model is trained for two additional epochs using three different methods. The comparison of their performance and additional training parameters are detailed in the appendix B."

4.1.2 Baselines

For the controlled sentiment generation task, we compare our method against ULMA and KTO. Due to the high annotation costs associated with DOVE, we defer this comparison to future work.

KTO requires adjustments to the parameters λ_D and λ_U based on the ratio of positive to negative examples. To maintain consistency as recommended, we fix the ratio of $\frac{\lambda_D n_D}{\lambda_U n_U}$ to 1. For example, in a scenario where $\frac{n_D}{n_U} = \frac{1}{4}$indicating only 20% of the data is positivewe set $\lambda_D = 4$ and $\lambda_U = 1$.

For all methods, including Retrieval-DPO, we standardize β at 1 to ensure a fair comparison.

4.1.3 Evaluation method

For the controlled sentiment generation task, we utilize a pre-existing sentiment classifier (Hartmann et al., 2023) we mentioned above as our reward model. This model assesses whether each generated review exhibits positive sentiment. We then calculate the proportion of positive cases within the entire test set.

For the single-turn dialogue task, we employ the gpt-4-0125-preview model as our evaluator. This model is used to compare the quality of generated responses against the chosen responses from the test set. Due to resource constraints, we randomly select 1,000 examples from the test set for evaluation in our experiments.

In both tasks, we use greedy sampling to generate the reviews or responses for further evaluation.

4.1.4 Computation environment

All the experiments in this paper were conducted on 4× A100 GPUs and the implementations are borrowed from TRL library[1]. The specific packages and codes used in our experiments will be available on our GitHub page.

4.2 Results and analysis

4.2.1 Performance analysis

The performance results are presented in Tables 1 and 2.

[1]https://huggingface.co/docs/trl/index

Table 1: Positive sentiment generation rate on controlled sentiment generation task. The first column for $A\%$ it means for $A\%$ of all the preference data, we only use the chosen response for training, and for other preference samples, we only use the rejected responses. The values mean the positive rate (%)

Percentage of positive examples	Retrieval-DPO	KTO	ULMA
20%	**99.22**	97.46	80.50
50%	**99.37**	97.55	80.85
80%	**99.34**	97.33	81.12

Table 2: Win/Tie rate over chosen responses on helpful-base subset and harmless-base subset of HH dataset. The first column for $A\%$ it means for $A\%$ of all the preference data, we only use the chosen response for training, and for other preference samples, we only use the rejected responses.

Percentage of positive examples	HH-helpful-base			HH-harmless-base		
	Retrieval-DPO	KTO	ULMA	Retrieval-DPO	KTO	ULMA
20%	**86.6/1.4**	73.5/4.2	70.9/1.8	74.6/2.2	75.2/2.3	X
50%	**86.3/1.8**	71.0/3.2	68.7/1.3	75.7/2.2	67.8/4.7	X
80%	**86.4/1.4**	67.9/3.5	69.5/1.2	76.0/1.8	75.0/2.4	X

Controlled sentiment generation task In this relatively straightforward task, ULMA achieves over 80% positive sentiment generation, and KTO surpasses 97%. Our method, Retrieval-DPO, performs even better than KTO, with nearly a 2% improvement, achieving more than 99% positive generation. This demonstrates a significant performance advantage of our method.

Single-Turn Dialogue Task (Helpful-Base Subset) Here, Retrieval-DPO also exhibits strong performance, maintaining approximately an 86% win rate. It demonstrates a substantial performance advantage of over 13% compared to the other baselines even if we consider the least gap (20% positive examples in the training set with KTO). This significant margin highlights the effectiveness of Retrieval-DPO in handling non-paired preference data.

Single-Turn Dialogue Task (Harmless-Base Subset) For this subset, ULMA-generated responses were frequently nonsensical, leading us to exclude it from this part of the evaluation. Both Retrieval-DPO and KTO generate normal responses, with Retrieval-DPO showing a slight performance advantage over KTO. This advantage may be attributed to the quality of the dataset.

4.2.2 Robustness to the ratio of positive examples to negative examples

From Table 1 and Table 2, we observe that Retrieval-DPO is far more robust to the ratio of positive examples to negative examples than KTO even we fixes $\frac{\lambda_D n_D}{\lambda_U n_U}$ to 1, as recommended in KTO paper. This phenomenon suggests that simply keeping this ratio in the range of $[1, \frac{4}{3}]$ might not be suffice for optimal performance. Compared to ULMA, our method is still more robust although both methods do need extra parameter tuning.

In contrast, Retrieval-DPO maintains a consistently high win rate of around 86% across all three percentages of positive examples in the HH-helpful-base subset. This consistency underscores that Retrieval-DPO's performance is less susceptible to fluctuations in the proportion of positive examples, evidencing its robustness. When compared to KTO and ULMA, Retrieval-DPO not only outperforms both in win rates across the HH dataset's subsets but also exhibits less performance variability as the percentage of chosen samples changes.

These observations reinforce the robustness of Retrieval-DPO, suggesting it as a more reliable choice in environments with varying data distributions.

Table 3: Win/Tie rate (%) over chosen responses on helpful-base subset of HH dataset

Percentage of positive examples	Random-DPO	Retrieval-DPO
20%	69.9/1.5	86.6/1.4
50%	70.1/1.4	86.3/1.8
80%	71.8/1.6	86.4/1.4

5 Further discussion

5.1 Is retrieval really helpful?

DOVE shows that randomly constructing paired preference data with annotation can perform well. Does our method's great performance only come from the construction of paired preference data rather than retrieving similar complementary data? We randomly select complementary data instead of retrieving and check its performance on the helpful-base of HH dataset. The results are shown in table 3 and we paste the results of Retrieval-DPO for comparison.

The table presents the win/tie rates (expressed as percentages) of two different methods, Random-DPO and Retrieval-DPO, on the helpful-base subset of the HH dataset. The data is categorized based on the percentage of positive examples: 20%, 50%, and 80%. Both Random-DPO and Retrieval-DPO show a win rate above 69% across all categories. However, Retrieval-DPO significantly outperforms Random-DPO in all cases, with a win rate consistently above 86%. This suggests that Retrieval-DPO is a more effective method overall. It supports our opinion that retrieving similar questions is necessary to keep high performance.

5.2 Adapting Retrieval to KTO

From equation 4, KTO assumes that the reference point z_0 should be related to all possible input-output pairs rather than one data point without more evidence support. We have already shown that retrieval can lead to giant performance gain when applied on DOVE. A natural question is can retrieval be adapted to other existing non-paired preference optimization algorithms? We only consider KTO as z_0 is an biased estimate thus might be inaccurate. Here we simply replace z_0 with $z_{retreival} = r_\theta(x_{retrieval}, y_{retrieval})$ and do not backpropagate through $z_{retrieval}$ as done in KTO. The results are shown in table 4.

From the results, replacing z_0 with retrieval cannot guarantee the better performance for KTO. One possible reason is that when the ratio is away from 1:1, the estimate of z_0 is bad compared to $z_retrieval$. Besides, even considering the best case, Retrieval-KTO's performance is still around 10 percentage points lower than that of Retrieval-DPO, showing that DOVE might be more suitable to KTO as a base optimization algorithm for retrieval.

6 Related work

Alignment Alignment is an important part for the modern LLMs' training, since the biases (Shah et al., 2019), safety problems (Gehman et al., 2020) and privacy issues (Carlini et al., 2021) in the training data have been found huge influence to the response of LLMs and bad responses really harm people's feeling. Hence we need to make the LLMs' responses be consistent with human preferences. Reinforcement Learning from Human Feedback (Ouyang et al., 2022) first introduced reinforcement learning into alignment and show strong performance in this field. In the training of RLHF, researchers need to first train a reward model then apply proximal policy optimization (Schulman et al., 2017) to increase the output rewards of the policy model. To avoid the high computing consumption, RRHF (Yuan et al., 2023) created a new loss function to modify the generating probability of the ranked responses, and DPO (Rafailov et al., 2023) showed a way to regard the language model as the reward model and align the language model with the human preference directly.

Non-paired Cai et al. (2023) propose point-wise DPO by separating desirable and undesirable examples in the DPO loss function and further propose ULMA loss as sum of the Supervised Fine-Tuning (SFT) loss for the desirable examples and an additional KL regularizer for the negative samples. Ethayarajh et al. (2024) show that Kahneman & Tverskys prospect theory can be be adapted to non-paired preference optimization problems and achieve comparable results with DPO even on

Table 4: Win/Tie rate (%) over chosen responses on helpful-base subset of HH dataset

Percentage of positive examples	Retrieval-KTO	KTO	Retrieval-DPO
20%	74.6/3.0	73.5/4.2	86.6/1.4
50%	69.0/3.3	71.0/3.2	86.3/1.8
80%	69.3/3.3	67.9/3.5	86.4/1.4

paired preference data.

Chen et al. (2024) propose Self-Play Fine-Tuning (SPIN) which generates the rejected responses from previous iterations. Yin et al. (2024) propose Self-Play Preference Optimization (SPPO) which dynamically adjusts training data in real-time instead of using pre-generated responses.

7 Future work

In the retrieval process, different retrievers will lead to different retrieval results. Therefore, for specific tasks, it is worth considering the work of training specific retrievers and designing specialized retrieval methods. Our current approach is to consider retrieving the optimal matching options. For the robustness of the model, we can consider using a Top-N retrieval method, selecting the best several examples as corresponding samples based on the current sample.

8 Conclusion

In this paper, we proposed a novel approach, Retrieval-DPO, for aligning Large Language Models (LLMs) with binary preference data. We leveraged retrieval methods to form preference pairs, which allows us to apply joint preference optimization techniques. Our method not only bypasses the need for costly paired preference data but also mitigates the complexity and instability associated with reinforcement learning approaches.

Our experimental results demonstrated that Retrieval-DPO outperforms other non-paired preference optimization algorithms such as Kahneman-Tversky Optimization (KTO) and Unified Language Model Alignment (ULMA) on two subsets of the HH dataset. Furthermore, we found that our approach has a low sensitivity to the ratio of desirable to undesirable data, which is a significant advantage when working with imbalanced datasets.

In conclusion, our work introduces an effective and efficient method for aligning LLMs with binary preference data, providing a promising direction for future research in this area. As we continue to refine and expand upon this method, we anticipate further improvements in the performance of LLMs in various real-world applications.

9 Limitations

While we conduct a comprehensive analysis of various non-paired preference optimization algorithms, we still miss some methods such as DOVE and SPIN as interesting baselines due to limited resources. We also leave more detailed analysis of how improvements of the retrieval module affect the pipeline as future work. For instance, we may unfreeze the retrieval model during the training phase or use a higher quality retrieval source and check whether it could lead to better performance of the pipeline. Although we tried to perform human evaluation to test whether it is decent to use GPT-4 as an evaluator, the questions in the helpful-base subset of HH are really hard to evaluate even for PhD candidates so we will try to find experts for the human evaluation when resources are available.

10 Potential Risks

In the retrieval process, we can retrieve positive and negative samples by accessing external knowledge bases. This approach requires ensuring the privacy and legality of the data, as well as guaranteeing that there are no data leakage issues in the retrieval database, in order to ensure the validity and security of our method.

References

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Hritik Bansal, Ashima Suvarna, Gantavya Bhatt, Nanyun Peng, Kai-Wei Chang, and Aditya Grover. 2024. Comparing bad apples to good oranges: Aligning large language models via joint preference optimization. *arXiv preprint arXiv:2404.00530*.

Tianchi Cai, Xierui Song, Jiyan Jiang, Fei Teng, Jinjie Gu, and Guannan Zhang. 2023. Ulma: Unified language model alignment with demonstration and point-wise human preference. *arXiv preprint arXiv:2312.02554*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Deven Shah, H Andrew Schwartz, and Dirk Hovy. 2019. Predictive biases in natural language processing models: A conceptual framework and overview. *arXiv preprint arXiv:1912.11078*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yueqin Yin, Zhendong Wang, Yujia Xie, Weizhu Chen, and Mingyuan Zhou. 2024. Self-augmented preference optimization: Off-policy paradigms for language model alignment. *arXiv preprint arXiv:2405.20830*.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*.

A Dataset examples

We exhibit some examples of our training datasets in table 5 and table 6

B Training parameters

Some training parameters for Retrieval-DPO, KTO and ULMA are listed in table 7.

C GPT-4 prompts for evaluating helpful and harmless on the single-turn dialogue task

Since our evaluation on the single-turn dialogue task relies on GPT-4, the prompts used for evaluation are important. We list our evaluation prompts here and they are adapted from DPO's evaluation prompts. Here we also consider the tie case although it is rare when GPT-4 acts as the evaluator. The order of generated and chosen responses are random for every evaluation.

For harmless-base dataset, to avoid the evaluator giving high preference to totally meaningless responses such as "I don't know.", we hope the evaluator to concern about the helpfulness as well.

Helpful-base:

```
For the following query to a chatbot, which response is more helpful?

Query: <the user query>

Response A:
<either the test method or baseline>

Response B:
<the other response>

FIRST provide a one-sentence comparison of the two responses and explain which you feel is more helpfu
Comparison: <one-sentence comparison and explanation>
More helpful: <"A" or "B" or "C">
```

Harmless-base:

```
For the following query to a chatbot, which response is more harmless while is still helpful to the qu

Query: <the user query>

Response A:
<either the test method or baseline>

Response B:
<the other response>

FIRST provide a one-sentence comparison of the two responses and explain which you feel is more harmle
Comparison: <one-sentence comparison and explanation>
```

```
More harmless while still helpful: <"A" or "B" or "C">
```

Table 5: IMDB-Prefix8 generated paired preference examples

Prompt	**Chosen**	**Rejected**
I love sci-fi and am willing	I love sci-fi and am willing to listen to what Iḿ about to read as long as it sounds good. This one is an exception. Thereś a lot of bad poetry, with some oddities in speech (such as that "What happened?" line at the middle of the story), but this movie has not got any such weaknesses. Even better than this, they managed to use CGI which gives it all a surreal aspect; one could call it post apocalyptic or science fiction or whatever you like, but therel̍l be no confusion nor are there really any obvious plot lines - just something that happens randomly and youŕe off doing your thing without thinking much about what goes on around you; we get time travel (if we can believe in that kind of crap) with aliens from far away...and then maybe another sequel...\n#9	I love sci-fi and am willing to forgive the low points here, but if you expect deep characters... sorry no.The plot isnt́ bad, just average Sci Fi. The romance scene has a lot of potential. So where are the interesting characters? How about that old guy with his strange glasses??? Thereś none! Just a generic human!I guess he was like that until some one introduced him who had better ideas which led to him becoming friends with young man named David, at least thats what I think they were planning in story telling way before he ended up getting hurt or died lol."Just when you thought she could not get any worse... SHE DID!"
No, no, no, no,	No, no, no, very. Very excellent. Just absolutely terrific! The dialogue at times is incredibly cheesy and has really nothing to do with the story and it's really annoying that those types of plot devices are so common in today's movies (and TV shows)! The story itself is fantastic! It's fun at the beginning but then things start getting even crazier as you get into this world which was not only quite strange but also completely off. I never saw anything similar to this happening before with these characters though so I didn't think much of them until recently Also if you don't like sci fi/fantasy type stuff why watch it? This isn't about a space battle either! It was like some sort of alien invasion set in modern day... or something haha -_-	No, no, no, no! And how can it be? Who do you think makes up the author/creator and the scriptwriter? When would you expect to see a title like this??? Why was she only allowed five episodes of that movie?? I know I'm being picky but honestly...the story in these four episodes is so bad and contrived. One thing I really liked about The Phantom Menace was that there were many twists during its run. Also it wasn't all trashy romantic comedies which are what i saw in this show!!! So please avoid this one!!

Table 6: hh-helpful paired preference examples

Prompt	**Chosen**	**Rejected**
Human: What are the rules to basketball? Assistant:	Well, its played by two teams of five people each. Each team tries to get the ball through the other teams hoop. The team that scores more points than their opponent is the winner.	The goal is to shoot the basketball through the hoop, in the most skillful way.
Human: Have any new technological advances been made in regards to electricity within the past few years?Assistant:	One thing thats recently gained a lot of attention is the potential for electricity to help with bone repair and regeneration. Some research suggests that electricity could be used as an artificial stimulus, essentially mimicking the action of biological bone growth and bone formation. This research is still in its early stages and requires further study, but could be especially interesting in the context of bone loss and treatment, in conjunction with potential new treatments for Osteoporosis. You may also find this video about the research interesting: https://www.youtube.com/watch?v=1r6U39KAcbQ	Let me think... oh yes, there have been some!

Table 7: Training parameters for controlled sentiment generation task and single-turn dialogue task when comparing all the methods

Parameter	Controlled sentiment generation	Single-turn dialogue
training-epochs	2	2
learning rate%	1e-6	1e-6
batch size	16	8
β	0.1	0.1