
LLM Alignment as Retriever Optimization: An Information Retrieval Perspective

Bowen Jin^{1,2} Jinsung Yoon² Zhen Qin³ Ziqi Wang¹ Wei Xiong¹ Yu Meng⁴ Jiawei Han¹ Sercan Ö. Arık²

Abstract

Large Language Models (LLMs) have revolutionized artificial intelligence with capabilities in reasoning, coding, and communication, driving innovation across industries. Their true potential depends on effective alignment to ensure correct, trustworthy and ethical behavior, addressing challenges like misinformation, hallucinations, bias and misuse. While existing Reinforcement Learning (RL)-based alignment methods are notoriously complex, direct optimization approaches offer a simpler alternative. In this work, we introduce a novel direct optimization approach for LLM alignment by drawing on established Information Retrieval (IR) principles. We present a systematic framework that bridges LLM alignment and IR methodologies, mapping LLM generation and reward models to IR’s retriever-reranker paradigm. Building on this foundation, we propose **LLM Alignment as Retriever Preference Optimization (LARPO)**, a new alignment method that enhances overall alignment quality. Extensive experiments validate LARPO’s effectiveness with 38.9 % and 13.7 % averaged improvement on AlpacaEval2 and MixEval-Hard respectively. Our work opens new avenues for advancing LLM alignment by integrating IR foundations, offering a promising direction for future research.

1. Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Team et al., 2024a) have demonstrated remarkable capacities in a wide range of fields including conversational modeling (Zhao et al., 2023a), reasoning (Wei et al., 2022) and code generation (Jiang et al., 2024). Unlocking the full poten-

tial of LLMs while ensuring their ethical, safe, and high-quality performance hinges on effective alignment (Wang et al., 2023). However, existing reinforcement learning-based LLM alignment methods (e.g., PPO (Ouyang et al., 2022)) involve multi-stage training and are challenging to optimize. To this end, direct LLM preference optimization methods (e.g., DPO (Rafailov et al., 2024)) are proposed to simplify the alignment process.

In this work, we further enhance direct LLM preference optimization, focusing on bringing Information Retrieval (IR) perspectives (Tay et al., 2022). Striking parallels exist between IR methodologies and LLM alignment techniques (Lin et al., 2022). For example, IR’s retriever-reranker framework, which uses a retriever for broad semantic matching to generate a candidate set and a reranker for fine-grained refinement, offers a compelling analogy to the Best-of-N approach in LLM alignment (Dong et al., 2023; Sessa et al., 2024). In this analogy, the LLM acts as the retriever, while the reward model serves as the reranker. Furthermore, the common use of dual-encoder architectures in both LLM generation and IR retrievers, coupled with the reliance on cross-encoder architectures in reward models and IR rerankers, further underscores this synergy. Leveraging established IR techniques offers the potential to develop novel, easily implementable LLM alignment methods grounded in IR principles, leading to improved alignment quality.

Despite the promising connections between LLM alignment and IR, a systematic exploration of this synergy remains lacking. Specifically, three key gaps exist: (1) a clear mapping between LLM alignment mechanisms and core IR principles has not been established; (2) empirical evaluations of LLMs through an IR lens are scarce; and (3) proven IR techniques like retriever optimization, hard negative mining, and candidate list construction are underutilized for LLM alignment. This paper directly addresses these gaps by systematically bridging LLM alignment and IR methodologies. Our contributions are fourfold:

- We introduce a comprehensive framework that connects LLM alignment techniques with the established IR principles, providing a new perspective on LLM alignment.
- We demonstrate the significance of three key IR prin-

¹University of Illinois at Urbana-Champaign ²Google Cloud AI Research ³Google DeepMind ⁴University of Virginia. Correspondence to: Bowen Jin <bowenj4@illinois.edu>.

ciples - retriever optimization objectives, hard negative mining, and candidate list construction - for improving LLM alignment.

- Building on these insights, we propose a novel alignment method, **LLM Alignment as Retriever Preference Optimization (LARPO)**, which demonstrably enhances alignment quality, with 38.9 % and 13.7 % relative averaged improvement on AlpacaEval2 and MixEval-Hard.
- We conduct further empirical studies to evaluate LLM performance using IR metrics, analyzing the impact of various post-training techniques.

In summary, this work establishes a crucial link between IR and LLM alignment, offering both novel insights and practical methods for advancing the field.

2. An Information Retrieval Perspective on LLMs

2.1. Primer on information retrieval

Information retrieval systems (Zhu et al., 2023) typically employ a two-stage process involving retrievers (Zhao et al., 2024) and rerankers (Lin et al., 2022). The retriever, often implemented as a bi-encoder (Figure 1), efficiently identifies a large set of (K) potentially relevant passages, denoted as $D_{\text{retrieval}}$, from a corpora C given a query q . This is achieved using a coarse-grained similarity function, $p_{\text{retrieval}}(d|q) = \text{Enc}_q^T(q) \cdot \text{Enc}_d(d)$, where Enc_q and Enc_d represent the query and passage encoders respectively:

$$D_{\text{retrieval}}(q) = \{d \in C \mid \max_{\text{top-}K} p_{\text{retrieval}}(d|q)\}. \quad (1)$$

However, due to the scale of the corpus, retrievers might not accurately capture fine-grained query-passage similarity with the simple dot production interaction function. Therefore, rerankers, typically implemented with cross-encoder (Figure 1), are employed to refine the ranking of the retrieved passages $D_{\text{retrieval}}$. The reranker produces a smaller set (k) of top-ranked passages, D_{rank} , using a fine-grained similarity function, $r_{\text{rank}}(q, d) = w \cdot \text{Enc}(q, d)$, where w is a learnable linear layer. Here, reranker adopts cross-encoder with both query/passage as inputs and encoded together while retriever adopts dual encoder for separate query/passage encoding.

$$D_{\text{rank}}(q) = \{d \in D_{\text{retrieval}}(q) \mid \max_{\text{top-}k} r_{\text{rank}}(q, d)\}. \quad (2)$$

The resulting ranked passages are ordered such that $D_{\text{rank}}(q) = \{d_1, d_2, \dots, d_k\}$ where $r_{\text{rank}}(q, d_1) \geq r_{\text{rank}}(q, d_2) \geq \dots \geq r_{\text{rank}}(q, d_k)$.

2.2. LLMs as retrievers. Reward models as rerankers

During inference, an LLM generates a response y given an input prompt x by modeling the probability distribution $p_{\text{LLM}}(y|x)$. Assuming a fixed maximum sequence length L and a vocabulary space V (Li et al., 2024), the set of all possible responses can be defined as $Y = \{y : y(1)y(2)\dots y(L) \mid y(i) \in V\} \subseteq V^L$.

We can conceptualize this process through an IR lens (Tay et al., 2022). The prompt x can be viewed as analogous to a query q , the set of all possible responses Y can be treated as the corpus C , and the generated response y can be considered as the retrieved passage d . Thus, given a prompt x , the LLM effectively acts as a retriever, searching for the most probable responses $D_{\text{LLM}}(x)$ from response space Y :

$$D_{\text{LLM}}(x) = \{y \in Y \mid \max_{\text{top-}K} p_{\text{LLM}}(d|x)\}. \quad (3)$$

where $p_{\text{LLM}}(y|x)$ is analogous to $p_{\text{retrieval}}(d|q)$ in IR.

This analogy is further supported by the LLMs' architecture. As illustrated in Figure 1, the generative modeling with LLMs can be interpreted as the matching process of a bi-encoder model. The prompt is encoded into a vector representation by LLM, while response tokens are represented as token embedding vectors. For each token position decoding, prompt embedding (obtained often from the hidden state of the last layer of the LLM) and vocabulary token embeddings are compared with a dot product, to determine the likelihood of a selected token for the response.

Furthermore, reward models $r_{\text{rm}}(x, y)$ (Lambert et al., 2024), which take both the prompt and response as input, function similarly to cross-encoders (*i.e.*, rerankers $r_{\text{rank}}(q, d)$ (Zhuang et al., 2023)) in IR. To enhance LLM performance, various inference-time strategies have been developed, including Best-of-N sampling (Stiennon et al., 2020) and majority voting (Wang et al., 2022). These can be interpreted as different configurations of retrievers and rerankers, as summarized in Appendix Table 5.

2.3. LLM tuning as retriever optimization

Supervised fine-tuning as direct retriever optimization.

Retriever training, aiming for accurate retrieval, often employs contrastive learning with the InfoNCE loss (Oord et al., 2018) to maximize $P(d_{\text{gold}}|q)$ of retrieving the ground truth passage d_{gold} given a query q . This can be expressed as:

$$\max \log P(d_{\text{gold}}|q) = \max \log \frac{\text{Enc}_d(d_{\text{gold}}) \cdot \text{Enc}_q(q)}{\sum_{j=1}^{|C|} \text{Enc}_d(d_j) \cdot \text{Enc}_q(q)}.$$

In the context of LLM alignment, supervised fine-tuning (SFT) aims to quickly adapt the model to a target task using prompt-response pairs (x, y_{gold}) . SFT maximizes the

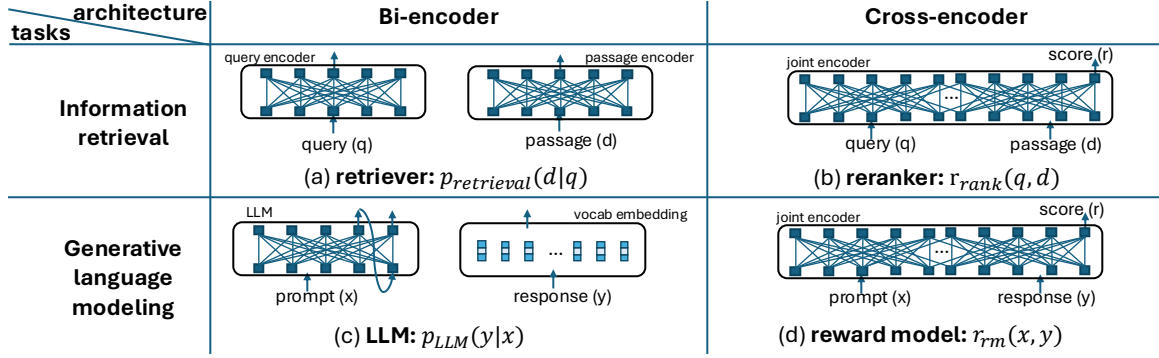


Figure 1. Architecture connection between retriever/LLM (bi-encoder) and reranker/reward model (cross-encoder). Bi-encoder models process each query/prompt and passage/response separately and often calculate their alignment score via a dot product operator, while cross-encoder models take both query/prompt and passage/response as input and score them directly. Bi-encoder models can be more efficient (*i.e.*, large-scale text matching) but the interaction between the two information unit is only captured by a dot production operation where their effectiveness can be constrained. Cross-encoder models can be more effective (*i.e.*, deeper interaction calculation with transformer architecture (Vaswani, 2017)) but less efficient. Although LLM involves auto-regressive token matching, which is different from retriever, some insights from IR can be borrowed to enhance LLM alignment as shown in the following sections.

conditional probability $P(y_{\text{gold}}|x)$ as:

$$\begin{aligned} \max \log P(y_{\text{gold}}|x) &= \max \log \prod_i^{y_{\text{gold}}} P(y_{\text{gold}}(i)|z_i) \\ &= \max \sum_i^{y_{\text{gold}}} \log \frac{\text{Emb}(y_{\text{gold}}(i)) \cdot \text{LLM}(z_i)}{\sum_{j=1}^{|V|} \text{Emb}(v_j) \cdot \text{LLM}(z_i)}, \end{aligned}$$

where $y(i)$ is the i -th token of y , $z_i = [x, y_{\text{gold}}(1 : i - 1)]$ represent the concatenation of the prompt x and the preceding tokens of y_{gold} , $\text{LLM}(\cdot)$ produces a contextualized representation, and $\text{Emb}(\cdot)$ is the token embedding function. We assume vocab embeddings and LLM hidden states share the same dimension, as in most LLMs.

Consequently, the SFT objective can be interpreted as a composite of multiple retrieval optimization objectives. In this analogy, $\text{LLM}(\cdot)$ acts as the query encoder and $\text{Emb}(\cdot)$ serves as the passage (or, in this case, token) encoder.

Preference optimization as reranker-retriever distillation. In retriever training, optimizing solely based on query/ground-truth document pairs can be suboptimal, particularly when using in-batch negatives for efficiency. Performance can be enhanced by distilling knowledge from a more powerful reranker to retriever (Qu et al., 2020; Zeng et al., 2022). This distillation process can be represented as $f_{\text{rerank}}(\cdot) \xrightarrow{c} \text{data} \xrightarrow{g(\cdot)} f_{\text{retrieval}}(\cdot)$, where new data, generated by the reranker $f_{\text{rerank}}(\cdot)$ based on a rule c , is used to optimize the retriever $f_{\text{retrieval}}(\cdot)$ with an objective $g(\cdot)$.

Similarly, in LLM alignment, a preference alignment phase often follows supervised fine-tuning (SFT) to further enhance the model using an external reward model to absorb preferential supervision effectively. Methods like PPO

(Schulman et al., 2017) and iterative DPO (Guo et al., 2024) exemplify this approach. Here, the LLM (considered acting as the retriever) generates responses that are then scored by the reward model (considered acting as the reranker). These scores are used to create new training data, effectively performing distillation from the reward model into the LLM: $f_{\text{reward-model}}(\cdot) \xrightarrow{c} \text{data} \xrightarrow{g(\cdot)} f_{\text{LLM}}(\cdot)$. Thus, preference optimization can be viewed as a form of reranker-to-retriever distillation, analogous to the process used in traditional IR.

We conduct empirical studies to understand SFT and preference optimization from IR perspective in Appendix B and have further discussion in Appendices C and D.

2.4. Empirical insights into LLMs as IR models

Evaluating LLMs as retrievers. A common metric for evaluating retrievers is Recall@ N , which assesses whether the top- N retrieved passages include any relevant passages for a given query. In the context of LLMs, this translates to evaluating whether the top- N generated responses contain a suitable response to the prompt, analogous to Pass@ N (Chen et al., 2021).

To draw the empirical connection between LLM and retrievers, we conduct an experiment on the GSM8K dataset (Cobbe et al., 2021) using Mathstral-7b-it (Mistral AI, 2025) and an experiment on the NQ dataset (Kwiatkowski et al., 2019) using e5 retriever. Figure 2 illustrates that increasing N can contribute to improved performance for both retriever and LLM. Detailed analysis can be found in Appendix E.

Greedy decoding, equivalent to $N = 1$, is a prevalent LLM inference strategy. However, as shown in Figure 2(b), allowing multiple attempts ($N > 1$) can substantially improve the chance of producing a correct answer, suggesting that per-

Table 1. LLM alignment objectives of LARPO. In the table, $\gamma(y | x) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$. All the proofs can be found in Appendix F.

Method	Assumption of $r(x, y)$	Objective
DPO	$\Pr(y_w \geq y_l) = \sigma(r(x, y_w) - r(x, y_l))$	$\mathcal{L}_{\text{pair}} = -\mathbb{E} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w x)}{\pi_{\text{ref}}(y_w x)} - \beta \log \frac{\pi_\theta(y_l x)}{\pi_{\text{ref}}(y_l x)} \right) \right]$
LARPO (Contrastive)	$\Pr(y_w \geq y_l^{(1)}, \dots, y_w \geq y_l^{(m)}) = \text{softmax}(r(x, y_w))$	$\mathcal{L}_{\text{con}} = -\mathbb{E} \left[\log \frac{\exp(\gamma(y_w x))}{\exp(\gamma(y_w x)) + \sum_{i=1}^m \exp(\gamma(y_l^{(i)} x))} \right]$
LARPO (LambdaRank)	$\Pr(y_1 \geq \dots \geq y_m) = \prod_{1 \leq i < j \leq m} \sigma(r(x, y_i) - r(x, y_j))$	$\mathcal{L}_{\text{lamb}} = -\mathbb{E} \left[\sum_{1 \leq i < j \leq m} \log \sigma(\gamma(y_i x) - \gamma(y_j x)) \right]$
LARPO (ListMLE)	$\Pr(y_1 \geq \dots \geq y_m) = \prod_{i=1}^m \text{softmax}_i^m(r(x, y_i))$	$\mathcal{L}_{\text{lmle}} = -\mathbb{E} \left[\sum_{i=1}^m \log \frac{\exp(\gamma(y_i x))}{\exp(\gamma(y_i x)) + \sum_{j=i+1}^m \exp(\gamma(y_j x))} \right]$

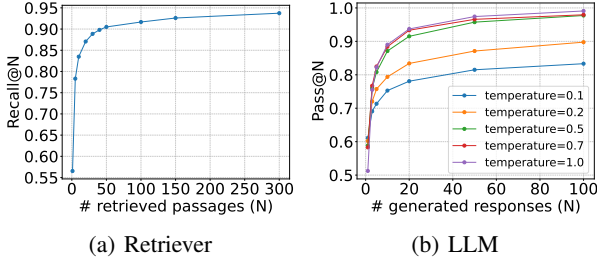


Figure 2. Analogy between evaluating retriever with Recall@N and LLM with Pass@N. As the number (N) of retrieved passages/generated responses increases, the retriever and LLM have a similar increasing trend. This highlights the importance of inference time scaling (e.g., Best-of-N) for LLM similar to retriever-reranker scaling in IR. Retriever: e5; LLM: Mathstral-7b-it.

formance under $N = 1$ may underestimate the model’s full potential. This highlights the importance of inference-time scaling techniques like Best-of-N (Stiennon et al., 2020) in LLM similar to retriever-reranker scaling (Zhuang et al., 2023) in IR. More results and analyses can be found in Appendix E.

3. Iterative LLM alignment as retriever optimization

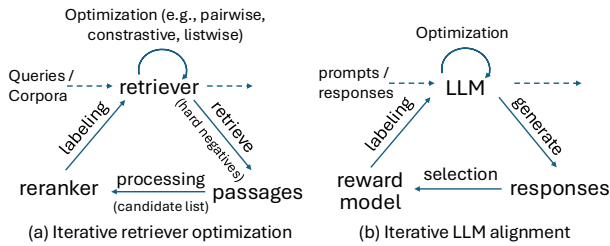


Figure 3. The connection between iterative LLM alignment (Xiong et al., 2024) and iterative retriever optimization (Xiong et al., 2020)

Iterative learning is a common technique in retriever optimization (Xiong et al., 2020), where results from the newly-trained model are used to generate new training data, as illustrated in Figure 3(a). Similarly, for LLM alignment, iterative preference optimization has been shown to enhance

performance (Xiong et al., 2024; Xu et al., 2024b; Guo et al., 2024) (Figure 3(b)). Drawing inspirations from retriever optimization, we re-examine iterative LLM preference optimization, focusing on three key aspects: (1) the optimization objective; (2) the use of hard negatives; and (3) the candidate list construction. Based on these aspects, we propose a new LLM alignment with an IR perspective, LARPO.

3.1. Retriever optimization objective

Typical objectives for retriever optimization include pairwise, contrastive and listwise objectives (Zhao et al., 2024). In this section, we discuss preference optimization variants (Wang et al., 2023) corresponding to different retriever optimization objectives. The optimization objective for preference optimization is given as:

$$\max_{\pi_{\text{LLM}}} \mathbb{E}_{x, y \sim \pi_{\text{LLM}}(\cdot|x)} [r(x, y)] - \beta \text{KL}(\pi_{\text{LLM}}(\cdot|x) || \pi_{\text{ref}}(\cdot|x)).$$

As discussed in Rafailov et al. (2024), the equation above has the optimal solution as:

$$r(x, y) = \beta \log \frac{\pi_{\text{LLM}}(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z, \quad (4)$$

where $Z = \sum_{y'} \pi_{\text{ref}}(y'|x) \exp(\frac{1}{\beta} r(x, y'))$ is the normalization constant and $r(\cdot)$ is the reward model which can also be seen as a reranker. According to different assumption for $r(x, y)$ from IR, we can obtain different training objectives as shown in Table 1, with proofs in Appendix F.

Pairwise ranking. Under the pairwise (Bradley-Terry) assumption $\Pr(y_w \geq y_l) = \sigma(r(x, y_w) - r(x, y_l))$, the policy objective becomes DPO (Rafailov et al., 2024) $\mathcal{L}_{\text{pair}}$.

Contrastive ranking. Another widely used objective for ranking is contrastive learning (Oord et al., 2018):

$$\Pr(y_w \geq y_l^{(1)}, \dots, y_w \geq y_l^{(m)}) = \text{softmax}(r(x, y_w)) = \frac{\exp(r(x, y_w))}{\exp(r(x, y_w)) + \sum_{i=1}^m \exp(r(x, y_l^{(i)}))}. \quad (5)$$

It handles multiple negatives in a single step, allowing the model to learn more robust representations for retrieval

and ranking. It is widely used for dense retriever training (Karpukhin et al., 2020). Under this ranking assumption, the policy objective becomes \mathcal{L}_{con} as shown in Table 1.

LambdaRank. In addition to pairwise and contrastive learning, list-wise ranking is widely adopted to sufficiently utilize the comprehensive information in candidate list. Inspired by LambdaRank (Burgess, 2010; Zeng et al., 2022):

$$\mathbb{P}\text{r}(y_1 \geq \dots \geq y_m) = \prod_{1 \leq i < j \leq m} \sigma(r(x, y_i) - r(x, y_j)), \quad (6)$$

the policy optimization objective becomes $\mathcal{L}_{\text{lamb}}$ (Table 1).

ListMLE. Another list-wise ranking assumption is the ListMLE assumption (Xia et al., 2008), which provides theoretical grounding and global optimization perspective:

$$\begin{aligned} \mathbb{P}\text{r}(y_1 \geq \dots \geq y_m) &= \prod_{i=1}^m \text{softmax}_i^m(r(x, y_i)) \\ &= \prod_{i=1}^m \frac{\exp(r(x, y_i))}{\exp(r(x, y_i)) + \sum_{j=i+1}^m \exp(r(x, y_j))} \end{aligned} \quad (7)$$

In this case, the objective becomes $\mathcal{L}_{\text{lmle}}$ shown in Table 1.

3.2. Hard negatives

Hard negatives are crucial for effective retriever training (Zhan et al., 2021; Qu et al., 2020), as learning to distinguish harder negatives potentially lead to more powerful retrievers (Xiong et al., 2020). In LLM alignment, negatives correspond to unpreferred responses (y_l) for a given prompt (x). In iterative on-policy training, various types of negatives can be identified, ordered by increasing difficulty: (1) **Easiest**: A random, unrelated response to x ; (2) **Easy**: A response to a related but different prompt (x'); (3) **Hard**: An incorrect response to x generated with a high temperature; (4) **Hardest**: An incorrect response to x generated with a low temperature.

Note that, assuming a well-initialized policy LLM, as indicated by Figure 2(b) ($N = 1$), low temperatures tend to produce harder negatives, yielding the above ranking. To be specific, lower temperatures yield more similar generated responses, increasing overlap between positive and negative samples. This effectively makes the negatives harder. According to Zhan et al. (2021), hardest negatives could be most important to LLM alignment.

3.3. Candidate list

In iterative retriever optimization, construction of the candidate list $[d_1, \dots, d_m]$, which is used by the reranker to generate data for the next iteration, is crucial. Prior research (Zeng et al., 2022) has identified factors such as list

Algorithm 1 LARPO: LLM alignment as iterative retriever preference optimization.

Require: Number of iterations T , number of new data per annotation phase M , number of generated responses for each prompt k , temperature for each iteration $\{t_i\}_{i=0}^T$, prompt dataset $\mathcal{D}_{\mathcal{X}} = \{x_i\}_{i=1}^N$, policy LLM π_{θ_0} , reward model r , learning rate γ , a ranking-based objective function $\mathcal{L}_{\text{rank}}$.

Ensure: Aligned LLM π_{θ_T} .

```

1: for  $s := 0$  to  $T$  do
2:   Update behavior LLM:  $\pi_{\beta} \leftarrow \pi_{\theta_s}$ 
3:   Preference dataset  $\mathcal{D}_s = \{\}$ 
4:   for  $i := 1$  to  $M$  do
5:     Sample prompt  $x \sim \mathcal{D}_{\mathcal{X}}$ 
6:     // candidate list construction
7:     Sample  $y_1, \dots, y_k \sim \pi_{\beta}(\cdot|x)_{t_s}$ 
8:     // hard negatives
9:     Rank  $\{y_i\}$  with  $r$ :  $Y_x = \{y_j^{(r)}\}$ , where  $(r(y_a^{(r)}) > r(y_b^{(r)}), a < b$ 
10:     $\mathcal{D}_s \leftarrow \mathcal{D}_s \cup \{(x, Y_x)\}$ 
11:   end for
12:   // candidate list construction
13:    $\mathcal{D} \leftarrow \text{Merge}_{k=0}^s \mathcal{D}_k$ 
14:   while  $\mathcal{D} \neq \emptyset$  do
15:     Sample a batch  $(x, Y_x)$  from  $\mathcal{D}$ 
16:     Update  $\mathcal{D} \leftarrow \mathcal{D} \setminus \{(x, Y_x)\}$ 
17:     // retriever optimization objective
18:      $\theta_s \leftarrow \theta_s - \gamma \cdot \nabla_{\theta} \mathcal{L}_{\text{rank}}(x, Y_x, \pi_{\theta}; \pi_{\beta})$ 
19:   end while
20:    $\theta_{s+1} \leftarrow \theta_s$ 
21: end for
    
```

size and candidate selection as being particularly important. Similarly, in iterative preference optimization, construction of the candidate response list $Y = [y_1, \dots, y_m]$ is critical. We identify two key factors influencing the quality of Y : inclusiveness and memorization.

- (1) **Inclusiveness** (Qu et al., 2020) refers to the **size** of the response list Y . A larger Y potentially encompasses more information.
- (2) **Memorization** (Zeng et al., 2022) refers whether previously generated responses Y' are included in the current list Y to preserve past results.

Given their importance in IR (Qu et al., 2020; Zeng et al., 2022), the impact of these factors on LLM alignment, however, remains largely under-explored.

4. The Proposed Solution: LARPO

Motivated by iterative retriever optimization pipeline as

Table 2. Evaluations on AlpacaEval 2 and MixEval. LC WR and WR denote length-controlled win rate and win rate respectively. Offline baseline performances on AlpacaEval 2 are from Meng et al. (2024b). We use LLM-blender (Jiang et al., 2023b) as the reward model for a fair comparison with the baselines and also report the result with a stronger reward model FsfairX (Dong et al., 2024)

Model	Mistral-Base (7B)				Mistral-Instruct (7B)			
	Alpaca Eval 2		MixEval	MixEval-Hard	Alpaca Eval 2		MixEval	MixEval-Hard
	LC WR	WR	Score	Score	LC WR	WR	Score	Score
SFT	8.4	6.2	0.602	0.279	17.1	14.7	0.707	0.361
Reward model: LLM-Blender (Jiang et al., 2023b)								
RRHF	11.6	10.2	0.600	0.312	25.3	24.8	0.700	0.380
SLiC-HF	10.9	8.9	0.679	0.334	24.1	24.6	0.700	0.381
DPO	15.1	12.5	0.686	0.341	26.8	24.9	0.702	0.355
IPO	11.8	9.4	0.673	0.326	20.3	20.3	0.695	0.376
CPO	9.8	8.9	0.632	0.307	23.8	28.8	0.699	0.405
KTO	13.1	9.1	0.704	0.351	24.5	23.6	0.692	0.358
RDPO	17.4	12.8	0.693	0.355	27.3	24.5	0.695	0.364
SimPO	21.5	20.8	0.672	0.347	32.1	34.8	0.702	0.363
Iterative DPO	18.9	16.7	0.660	0.341	20.4	24.8	0.719	0.389
LARPO (Contrastive)	31.6	30.8	0.703	0.409	32.7	38.6	0.718	0.418
LARPO (LambdaRank)	34.9	37.2	0.695	0.452	32.9	38.9	0.720	0.417
LARPO (ListMLE)	31.1	32.1	0.669	0.390	29.7	36.2	0.709	0.397
Reward model: FsfairX (Dong et al., 2024)								
LARPO (Contrastive)	41.5	42.9	0.718	0.417	43.0	53.8	0.718	0.425
LARPO (LambdaRank)	35.8	34.1	0.717	0.431	41.9	48.1	0.740	0.440
LARPO (ListMLE)	36.6	37.8	0.730	0.423	39.6	48.1	0.717	0.397

shown in Figure 3(a) and the three key points in IR, we introduce LARPO, a novel approach to LLM alignment formulated as iterative retriever preference optimization. The algorithmic details are provided in Algorithm 1. Specifically, our experimental setup explores the following key aspects: (1) **Optimization objective**: We evaluate three distinct loss functions as the ranking objective ($\mathcal{L}_{\text{rank}}$): \mathcal{L}_{con} , $\mathcal{L}_{\text{lamb}}$, and $\mathcal{L}_{\text{lmle}}$. (2) **Hard negatives**: For a given prompt, hard negative samples are constructed by selecting less preferred responses generated with an appropriate temperature through parameter search. More details of the temperature are available in Appendix H.1. (3) **Candidate list**: In each iteration, we generate multiple (10) candidate responses considering inclusiveness. In terms of memorization, the candidate pool for subsequent iterations includes all previously generated responses.

5. Main Results

Baselines. We evaluate the performance of LARPO against a range of established preference optimization methods, encompassing both offline and online approaches. Our offline comparison set includes RRHF (Yuan et al., 2023), SLiC-HF (Zhao et al., 2023b), DPO (Guo et al., 2024), IPO (Azar et al., 2024), CPO (Xu et al., 2024a), KTO (Ethayarajh et al., 2024), RDPO (Park et al., 2024) and SimPO (Meng et al., 2024b). For online methods, we compare with iterative DPO (Xiong et al., 2024). The baseline checkpoints

are from Meng et al. (2024b). Further details regarding these baselines and our experimental setup are provided in Appendix G. Both baselines and LARPO are trained on Ultrafeedback dataset (Cui et al., 2024) for fair comparison.

Datasets. We conduct evaluation on two widely used benchmarks AlpacaEval2 (Dubois et al., 2024) and MixEval (Ni et al., 2024). These benchmarks are designed to assess the conversational capabilities of models across a diverse range of queries. AlpacaEval2 comprises 805 questions sourced from five datasets, while MixEval includes 4000 general and 1000 hard questions. Evaluation follows the established protocols for each benchmark. For AlpacaEval 2, we report both the raw win rate (WR) and the length-controlled win rate (LC). These benchmarks collectively provide a comprehensive assessment of the models’ instruction-following and problem-solving capabilities.

Results. The baseline performances on AlpacaEval 2 are directly from Meng et al. (2024b), while the performances on MixEval is evaluated by ourselves with the opensourced checkpoints. We adopt the same LLM-Blender (Jiang et al., 2023b) reward model for a fair comparison with the baselines and also explore stronger reward model: FsfairX (Dong et al., 2024). The results, presented in Table 2, show that LARPO consistently outperforms the competitive baseline methods on both datasets, with 38.9 % and 13.7 % averaged relative improvements, on AlpacaEval2 and MixEval-

Hard respectively, with the same reward model as the baselines. With a stronger reward model, we can further improve LARPO by 25.8 % on the challenging AlpacaEval2 dataset. Additional details regarding our experimental setup are available in Appendix H.1.

6. Analyses

This section provides empirical analyses of the three factors identified in Section 3.

6.1. Retriever optimization objective

Experimental setting. Iterative preference optimization is performed on LLMs using the different learning objectives outlined in Section 3.1. Alignment experiments are conducted using the Gemma2-2b-it (Team et al., 2024b) and Mistral-7b-it (Jiang et al., 2023a) models, trained on the Ultrafeedback dataset (Cui et al., 2024). Following the methodology of (Dong et al., 2024), we conduct three iterations of training and report the performance of the final checkpoint in Table 3. Model evaluations are performed on AlpacaEval2 (Dubois et al., 2024) and MixEval (Ni et al., 2024). Detailed settings can be found in Appendix H.2.

Table 3. Preference optimization objective study on AlpacaEval2 and MixEval. SFT corresponds to the initial chat model.

Method	AlpacaEval 2		MixEval	MixEval-Hard
	LC Winrate	Winrate	Score	Score
Gemma2-2b-it	SFT	36.39	38.26	0.6545
	pairwise	41.39	54.60	0.6740
	contrastive	43.41	56.83	0.6745
	ListMLE	49.77	62.05	0.6715
	LambdaRank	43.76	60.56	0.6750
Mistral-7b-it	SFT	21.14	14.22	0.7070
	pairwise	36.43	41.86	0.7175
	contrastive	38.44	42.61	0.7260
	ListMLE	38.02	43.03	0.7360
	LambdaRank	40.29	46.21	0.7370

Observation. Table 3 presents the results, from which we make the following observations: (1) Contrastive optimization generally outperforms pairwise optimization (*e.g.*, DPO), likely due to its ability to incorporate more negative examples during each learning step. (2) Listwise optimization methods, including ListMLE and LambdaRank, generally demonstrate superior performance compared to both pairwise and contrastive approaches. This is attributed to their utilization of a more comprehensive set of preference information within the candidate list.

6.2. Hard negatives

Experimental setting. The Mathstral-7b-it model is trained on the GSM8k training set and evaluated its performance on the GSM8k test set. Iterative DPO is employed as the RLHF method, with the gold or correct response

designated as the positive example. The impact of different hard negative variants is investigated, as described in Section 3.2, with the results presented in Figure 4(a). Additionally, the influence of temperature on negative hardness with Lambdarank objective are examined using experiments on the AlpacaEval 2 dataset, with results shown in Figure 4(b). Detailed settings are in Appendix H.5 and H.6.

Observation. Figure 4(a) illustrates that the effectiveness of the final LLM is directly correlated with the hardness of the negatives used during training. Harder negatives consistently lead to a more performant LLM. Figure 4(b) further demonstrates that, within a specific range, lower temperatures generate harder negatives, resulting in a more effective final trained LLM. However, much lower temperature could lead to less diverse responses and finally lead to LLM alignment performance drop.

6.3. Candidate List

Experimental setting. To investigate the impact of inclusiveness and memorization on LLM alignment, experiments are conducted using Gemma2-2b-it, employing the same training settings as in our objective study. For the inclusiveness study, the performance of the trained LLM is evaluated using varying numbers of candidates in the list. For the memorization study, three approaches are compared: (i) using only the current iteration’s responses, (ii) using responses from the current and previous iteration, and (iii) using responses from the current and all previous iterations. Detailed settings can be found in Appendix H.7 and H.3.

Table 4. Candidate list study with $\mathcal{L}_{\text{pair}}$ on Gemma2-2b-it. Previous iteration responses enhance performance.

Method	Alpaca Eval 2	
	LC Winrate	Winrate
SFT	47.03	48.38
Alignment (w. current)	55.06	66.56
Alignment (w. current + prev)	55.62	70.92
Alignment (w. current + all prev)	56.02	72.50

Observation. Figure 4(c) illustrates the significant impact of candidate list size on LLM alignment performance. As the candidate list size increases, performance improves, albeit with a diminishing rate of return. This is intuitive, given that a bigger candidate list size can contribute to more hard negatives and potentially benefit the model learning (Qu et al., 2020). Table 4 demonstrates that incorporating responses from previous iterations can enhance performance. This is potentially because introducing previous responses can make the candidate list more comprehensive and lead to better preference signal capturing. More explanations are in Appendix H.3.

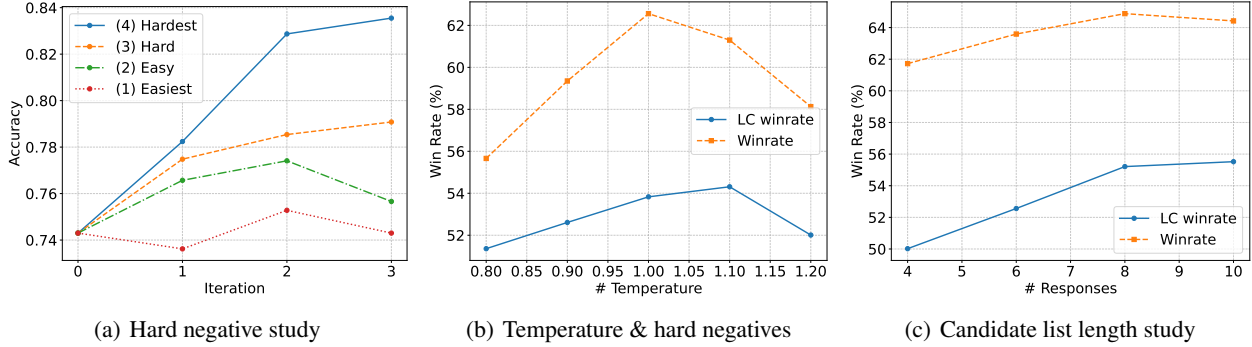


Figure 4. Hard negative and candidate list study. (a) Hard negative study with $\mathcal{L}_{\text{pair}}$ on GSM8K with Mathstral-7b-it model. We explore four negative settings: (1) a random response not related to the given prompt; (2) a response to a related prompt; (3) an incorrect response to the given prompt with high temperature; (4) an incorrect response to the given prompt with suitable temperature. Hardness: (4)>(3)>(2)>(1). The harder the negatives are, the stronger the trained LLM is. (b) Training temperature study with $\mathcal{L}_{\text{pair}}$ on Mistral-7b-it and Alpaca Eval 2. Within a specific range (> 1), lower temperature leads to harder negative and benefit the trained LLM. However, much lower temperature could lead to less diverse responses and finally lead to LLM alignment performance drop. (c) Candidate list size study with \mathcal{L}_{con} on Mistral-7b-it. As the candidate list size increases, alignment performance improves.

7. Related works

LLM alignment. Pretrained LLMs demonstrate remarkable capabilities across a broad spectrum of tasks (Brown et al., 2020). Their performance at downstream tasks, such as conversational modeling, is significantly enhanced through alignment with human preferences (Ouyang et al., 2022; Bai et al., 2022). RLHF (Christiano et al., 2017) has emerged as a foundational framework for this alignment, typically involving learning a reward function via a preference model, often using the Bradley-Terry model (Bradley & Terry, 1952), and tuning the LLM using reinforcement learning (RL) to optimize this reward. Despite its success, RLHF’s practical implementation is notoriously complex, requiring multiple LLMs, careful hyperparameter tuning, and navigating challenging optimization landscapes.

Recent research has focused on simplifying this process. A line of works studies the direct alignment algorithms (Zhao et al., 2023b; Rafailov et al., 2024; Azar et al., 2024), which directly optimize the LLM in a supervised manner without first constructing a separate reward model. In particular, the representative DPO (Rafailov et al., 2024) attracts significant attention in both academia and industry. After these, SimPO (Meng et al., 2024b) simplifies DPO by using length regularization in place of a reference model.

Although LLMs are adopted for IR (Tay et al., 2022), there is a lack of study to improve direct LLM alignment with IR principles. This paper fills this gap by establishing a systematic link between LLM alignment and IR methodologies, and introducing a novel iterative LLM alignment approach that leverages insights from retriever optimization to advance the state of the art. The most related work is LiPO (Liu et al., 2024), which applies learning-to-rank objectives. However, LiPO relies on off-the-shelf listwise preference

data, which is hard to satisfy in practice.

Language models for information retrieval. Language models (LMs) have become integral to modern IR systems (Zhu et al., 2023), particularly after the advent of pretrained models like BERT (Devlin, 2019). A typical IR pipeline employs retrievers and rerankers, often based on dual-encoder and cross-encoder architectures, respectively (Humeau, 2019). Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) pioneered the concept of dense retrieval, laying the groundwork for subsequent research. Building on DPR, studies have emphasized the importance of hard negatives in training (Zhan et al., 2021; Qu et al., 2020) and the benefits of online retriever optimization (Xiong et al., 2020).

In the realm of reranking, Nogueira & Cho (2019) were among the first to leverage pretrained language models for improved passage ranking. This was followed by MonoT5 (Nogueira et al., 2020), which scaled rerankers using large encoder-decoder transformer architectures, and RankT5 (Zhuang et al., 2023), which introduced pairwise and listwise ranking objectives. Recent work has also highlighted the importance of candidate list preprocessing before reranking (Meng et al., 2024a).

Despite the pervasive use of LMs in IR, the interplay between LLM alignment and IR paradigms remains largely unexplored. This work aims to bridge this gap, establishing a strong connection between LLM alignment and IR, and leveraging insights from both fields to advance our understanding of LLM alignment from an IR perspective.

8. Conclusions

This paper has forged a novel link between LLM alignment and IR, offering a systematic framework to enhance the LLM alignment performance. Expanding upon this basis, we introduced LARPO, a new direct preference optimization method that integrates the IR principles to significantly enhance alignment quality. The effectiveness of LARPO is strongly supported by our comprehensive experiments across widely-used benchmarks, demonstrating its potential as a significant advancement in LLM alignment. Furthermore, our IR-focused analysis highlights the crucial role of retriever optimization objectives, hard negatives, and candidate list construction in achieving effective alignment.

Acknowledgements

This research was supported in part by Apple PhD Fellowship, in part by US DARPA INCAS Program No. HR0011-21-C0165 and BRIES Program No. HR0011-24-3-0325, in part by the Office of Naval Research contract number N000142412612, in part by NSF grant numbers IIS-19-56151 and 2402873, in part by the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897 and the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329, in part by Cisco, and in part by the Center for Intelligent Information Retrieval. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of the sponsors or the U.S. Government.

Impact Statement

This paper contributes to the advancement of machine learning. While our work may have broader societal implications, we do not believe any specific impacts warrant explicit discussion in this context.

References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.
- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., Das-Sarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.
- Burges, C. J. From ranknet to lambdarank to lambdamart: An overview. *Learning*, 11(23-581):81, 2010.
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Cui, G., Yuan, L., Ding, N., Yao, G., He, B., Zhu, W., Ni, Y., Xie, G., Xie, R., Lin, Y., et al. Ultrafeedback: Boosting language models with scaled ai feedback. In *Forty-first International Conference on Machine Learning*, 2024.
- Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.
- Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., Shum, K., and Zhang, T. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.
- Dong, H., Xiong, W., Pang, B., Wang, H., Zhao, H., Zhou, Y., Jiang, N., Sahoo, D., Xiong, C., and Zhang, T. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.
- Dubois, Y., Galambosi, B., Liang, P., and Hashimoto, T. B. Length-controlled alpacaEval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.
- Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

- Guo, S., Zhang, B., Liu, T., Liu, T., Khalman, M., Llinares, F., Rame, A., Mesnard, T., Zhao, Y., Piot, B., et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.
- Humeau, S. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*, 2019.
- Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a.
- Jiang, D., Ren, X., and Lin, B. Y. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*, 2023b.
- Jiang, J., Wang, F., Shen, J., Kim, S., and Kim, S. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024.
- Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.
- Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.
- Lambert, N., Pyatkin, V., Morrison, J., Miranda, L., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.
- Li, X., Jin, J., Zhou, Y., Zhang, Y., Zhang, P., Zhu, Y., and Dou, Z. From matching to generation: A survey on generative information retrieval. *arXiv preprint arXiv:2404.14851*, 2024.
- Lin, J., Nogueira, R., and Yates, A. *Pretrained transformers for text ranking: Bert and beyond*. Springer Nature, 2022.
- Liu, T., Qin, Z., Wu, J., Shen, J., Khalman, M., Joshi, R., Zhao, Y., Saleh, M., Baumgartner, S., Liu, J., et al. Lipo: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878*, 2024.
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhunoye, S., Yang, Y., et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36, 2024.
- Meng, C., Arabzadeh, N., Askari, A., Aliannejadi, M., and de Rijke, M. Ranked list truncation for large language model-based re-ranking. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 141–151, 2024a.
- Meng, Y., Xia, M., and Chen, D. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024b.
- Mistral AI. Introducing mathstral, 2025. URL <https://mistral.ai/news/mathstral/>. Accessed: 2025-01-16.
- Ni, J., Xue, F., Yue, X., Deng, Y., Shah, M., Jain, K., Neubig, G., and You, Y. Mixeval: Deriving wisdom of the crowd from llm benchmark mixtures. *arXiv preprint arXiv:2406.06565*, 2024.
- Nogueira, R. and Cho, K. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.
- Nogueira, R., Jiang, Z., and Lin, J. Document ranking with a pretrained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*, 2020.
- Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Park, R., Rafailov, R., Ermon, S., and Finn, C. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.
- Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., Dong, D., Wu, H., and Wang, H. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2010.08191*, 2020.
- Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Sessa, P. G., Dadashi, R., Hussenot, L., Ferret, J., Vieillard, N., Ramé, A., Shariari, B., Perrin, S., Friesen, A.,

- Cideron, G., et al. Bond: Aligning llms with best-of-n distillation. *arXiv preprint arXiv:2407.14622*, 2024.
- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.
- Tay, Y., Tran, V., Dehghani, M., Ni, J., Bahri, D., Mehta, H., Qin, Z., Hui, K., Zhao, Z., Gupta, J., et al. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35: 21831–21843, 2022.
- Team, G., Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024a.
- Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024b.
- Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Huang, S., Rasul, K., Bartolome, A., M. Rush, A., and Wolf, T. The Alignment Handbook. URL <https://github.com/huggingface/alignment-handbook>.
- Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.
- Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., and Liu, Q. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*, 2023.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Xia, F., Liu, T.-Y., Wang, J., Zhang, W., and Li, H. List-wise approach to learning to rank: theory and algorithm. In *Proceedings of the 25th international conference on Machine learning*, pp. 1192–1199, 2008.
- Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P., Ahmed, J., and Overwijk, A. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*, 2020.
- Xiong, W., Dong, H., Ye, C., Wang, Z., Zhong, H., Ji, H., Jiang, N., and Zhang, T. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.
- Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Van Durme, B., Murray, K., and Kim, Y. J. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*, 2024a.
- Xu, W., Li, J., Wang, W. Y., and Li, L. Bpo: Staying close to the behavior llm creates better online llm alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11125–11139, 2024b.
- Yu, L., Jiang, W., Shi, H., Yu, J., Liu, Z., Zhang, Y., Kwok, J. T., Li, Z., Weller, A., and Liu, W. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.
- Yuan, Z., Yuan, H., Tan, C., Wang, W., Huang, S., and Huang, F. Rhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.
- Zeng, H., Zamani, H., and Vinay, V. Curriculum learning for dense retrieval distillation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1979–1983, 2022.
- Zhan, J., Mao, J., Liu, Y., Guo, J., Zhang, M., and Ma, S. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1503–1512, 2021.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023a.
- Zhao, W. X., Liu, J., Ren, R., and Wen, J.-R. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60, 2024.
- Zhao, Y., Joshi, R., Liu, T., Khalman, M., Saleh, M., and Liu, P. J. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023b.

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Liu, Z., Dou, Z., and Wen, J.-R. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*, 2023.

Zhuang, H., Qin, Z., Jagerman, R., Hui, K., Ma, J., Lu, J., Ni, J., Wang, X., and Bendersky, M. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2308–2313, 2023.

A. LLM inference strategy and IR pipelines

Table 5. Correspondence between LLM inference and IR pipelines.

Method	Retriever	Reranker	Pipeline
Greedy decoding	LLM	\emptyset	Retriever-only
Best-of-N (Stiennon et al., 2020)	LLM	Reward model	Retriever-reranker
Majority voting (Wang et al., 2022)	LLM	Majority	Retriever-reranker
Iterative refinement (Madaan et al., 2024)	LLM	\emptyset	Iterative retrieval w. query rewriting

B. How can SFT and preference optimization help the LLM from an IR perspective?

We assess how well LLMs perform at two tasks: fine-grained reranking (using greedy decoding accuracy) and coarse-grained retrieval (using Recall@N). We focus on how SFT and DPO, affect these abilities. Using the Mistral-7b model, we evaluate on the GSM8k and MATH datasets with two approaches: SFT-only, and SFT followed by DPO (SFT \rightarrow DPO).

In the SFT phase, the model is trained directly on correct answers. For DPO, we generate 20 responses per prompt and created preference pairs by randomly selecting one correct and one incorrect response. We use hyperparameter tuning and early stopping to find the best model checkpoints (see Appendix H.4 for details).

Table 6. Retrieval (Recall@N) and reranking (greedy accuracy) metrics across dataset and training strategies, with Mistral-7b as the LLM. 0.7 is used as the temperature. Recall@N can also be denoted as pass@N.

	Metric	init model	SFT	SFT \rightarrow DPO
GSM8K	Greedy Acc	0.4663	0.7680	0.7991
	Recall@20	0.8347	0.9462	0.9545
	Recall@50	0.9090	0.9629	0.9727
	Recall@100	0.9477	0.9735	0.9826
Math	Greedy Acc	0.1004	0.2334	0.2502
	Recall@20	0.2600	0.5340	0.5416
	Recall@50	0.3354	0.6190	0.6258
	Recall@100	0.4036	0.6780	0.6846

The results are shown in Table 6. We observe that both SFT and DPO improve both retrieval and reranking, with SFT being more effective. Adding DPO after SFT further improves performance on both tasks. This is consistent with information retrieval principles that both direct retriever optimization and reranker-retrieval distillation can enhance the retriever performance, while the latter on top of the former can further improve the performance. Further discussions can be found in Appendices C and D.

C. Discussion on the connection and difference between SFT and direct retriever optimization

As discussed in Section 2.3, the direct retriever optimization goal with InfoNCE is shown as:

$$\max \log P(d_{\text{gold}}|q) = \max \log \frac{\text{Enc}_d(d_{\text{gold}}) \cdot \text{Enc}_q(q)}{\sum_{j=1}^{|C|} \text{Enc}_d(d_j) \cdot \text{Enc}_q(q)},$$

while the SFT optimization goal is shown as:

$$\max \log P(y_{\text{gold}}|x) = \max \log \prod_i^{|y_{\text{gold}}|} P(y_{\text{gold}}(i)|z_i) = \max \sum_i^{|y_{\text{gold}}|} \log \frac{\text{Emb}(y_{\text{gold}}(i)) \cdot \text{LLM}(z_i)}{\sum_{j=1}^{|V|} \text{Emb}(v_j) \cdot \text{LLM}(z_i)}. \quad (8)$$

As a result, the SFT objective can be seen as a summation of multiple retrieval optimization objectives, where $\text{LLM}(\cdot)$ and word embedding $\text{Emb}(\cdot)$ are query encoder and passage encoder respectively.

However, for direct retriever optimization with InfoNCE, $\text{Enc}_d(\cdot)$ is usually a large-scale pretrained language model which is computationally expensive on both time and memory. In this case, it is unrealistic to calculate the $\text{Enc}_d(d_j)$ for all $d_j \in C$, when C is large, because of the time constrain and GPU memory constrain. As a result, a widely-adopted technique is to adopt “in-batch negatives” with “hard negatives” to estimate the $\log P(d_{\text{gold}}|q)$ function:

$$\max \log P(d_{\text{gold}}|q) = \max \log \frac{\text{Enc}_d(d_{\text{gold}}) \cdot \text{Enc}_q(q)}{\sum_{j=1}^{|C|} \text{Enc}_d(d_j) \cdot \text{Enc}_q(q)} \sim \max \log \frac{\text{Enc}_d(d_{\text{gold}}) \cdot \text{Enc}_q(q)}{\sum_{i=1}^{|B|} \text{Enc}_d(d_i) \cdot \text{Enc}_q(q) + \sum_{j=1}^{|H|} \text{Enc}_d(d_j) \cdot \text{Enc}_q(q)},$$

where B is the in-batch negative set and H is the hard negative set. Note that $B \cup H \subset C$. This objective is more efficient to optimize but is not the original optimization goal. As a result, the learned model after direct retriever optimization is not optimal. It is also found that the hard negatives H is the key to estimate the original optimization goal (Zhan et al., 2021). Thus, reranker-retriever distillation can further improve the retriever by introducing more hard negatives.

On the other hand, LLM optimization, as shown in Eq. (8), can be seen as a summation of multiple retrieval optimization function. In each retrieval step, the passage can be seen as a token and the corpus is the vocabulary space V . Given that the passage encoder $\text{Emb}(\cdot)$ (word embedding) here is cheap to compute and the vocabulary space V ($< 100k$) is usually not as large as C ($> 1M$) in IR, the objective in Eq. (8) can be directly optimized without any estimation. In this case, the LLM as a retriever is more sufficiently trained compared with the retriever training in IR.

D. Discussion on the connection and difference between preference optimization and reranker-retriever distillation

As discussed in Section 2.3, preference optimization with an online reward model $f_{\text{reward-model}}(\cdot) \xrightarrow{r} \text{data} \xrightarrow{g(\cdot)} f_{\text{LLM}}(\cdot)$ can be seen as a reranker to retriever distillation process $f_{\text{rerank}}(\cdot) \xrightarrow{r} \text{data} \xrightarrow{g(\cdot)} f_{\text{retrieval}}(\cdot)$, where the reward model is the reranker (*i.e.*, cross-encoder) and the LLM is the retriever (*i.e.*, bi-encoder).

However, there are two slight differences here:

- The LLM after SFT is more sufficiently trained compared to a retriever after direct optimization. As discussed in Appendix C, the SFT optimization function is not an estimated retriever optimization goal compared with the direct retrieval optimization. As a result, the LLM after SFT is sufficiently trained. In this case, if the reward model (reranker) cannot provide information other than that already in the SFT set (*e.g.*, using the SFT prompts), this step may not contribute to significant LLM capability improvement.
- The reward model may introduce auxiliary information than the reranker in IR. For a reranker in IR, it captures a same semantic with the retriever: semantic similarity between the query and the passage. However, in LLM post-training, the goal and data in SFT and preference optimization can be different. For example, the SFT phase could have query/response pairs which enable basic chat-based retrieval capability for the LLM. While the reward model may contain some style preference information or safety information which do not exist in SFT data. In this case, the preference optimization which is the reranker to retriever distillation step could also contribution to performance improvement.

E. Evaluate LLMs as retrievers

In addition to Mathstral-7b-it on GSM8K in Figure 2, we conduct extensive experiments to both Mistral-7b-it and Mathstral-7b-it on GSM8K and MATH. The results are shown in Figure 5. We have similar findings as in Figure 2 that: (1) As N increases, Recall@ N improves significantly, indicating that retrieving a larger number of documents increases the likelihood of including a correct one within the set. (2) For smaller values of N (*e.g.*, $N = 1$), lower temperatures yield higher Recall@ N . This is because lower temperatures reduce response randomness, favoring the selection of the most relevant result. (3) Conversely, for larger N (*e.g.*, $N > 10$), higher temperatures enhance Recall@ N . Increased temperature promotes greater response diversity, which, when combined with a larger retrieval set, improves the chances of capturing the correct answer within the results.

F. LARPO retriever optimization objective

We provide the proof for different variants of LARPO’s objective functions.

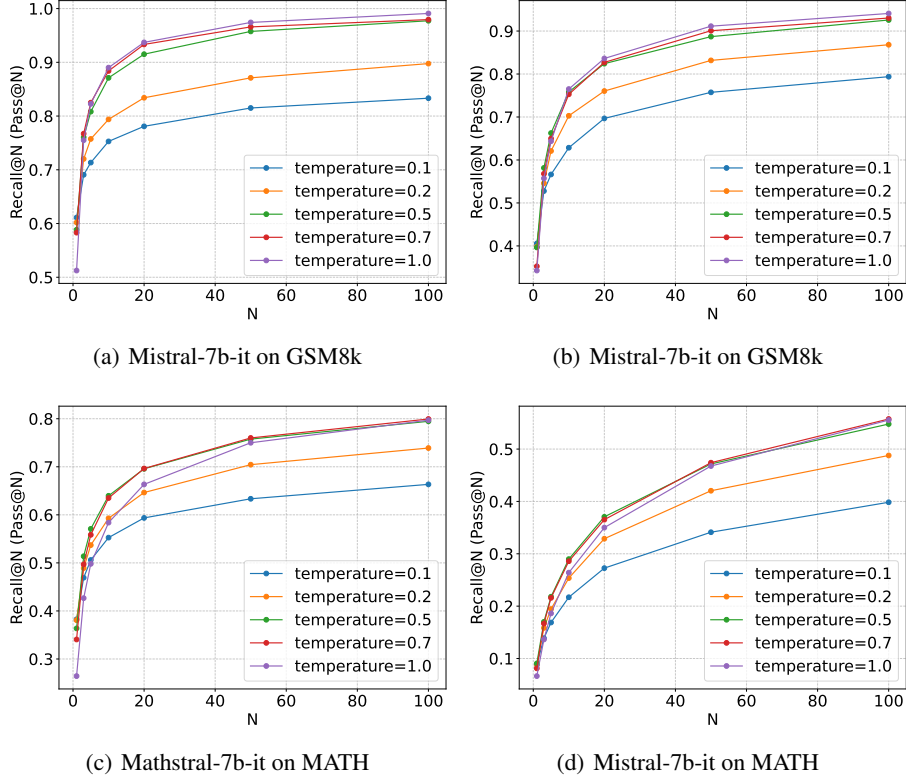


Figure 5. Evaluate the LLM as a retriever with Recall@N (Pass@N). As the number (N) of retrieved responses increases, the retrieval recall increases. The higher the temperature is, the broader spectrum the retrieved responses are, and thus the higher the recall is.

F.1. Contrastive ranking

Theorem F.1. Let x be a prompt and $(y_w, y_l^{(1)}, \dots, y_l^{(m)})$ be the responses for x under the contrastive assumption (Eq.(5)). Then the objective function to learn the LLM π_θ :

$$\mathcal{L}_{\text{con}} = -\mathbb{E} \left[\log \frac{\exp(\gamma(y_w | x))}{\exp(\gamma(y_w | x)) + \sum_{i=1}^m \exp(\gamma(y_l^{(i)} | x))} \right], \quad (9)$$

where $\gamma(y | x) = \beta \log \frac{\pi_\theta(y | x)}{\pi_{\text{ref}}(y | x)}.$

Proof. From (Rafailov et al., 2024), we know that

$$r(x, y) = \beta \log \frac{\pi_{\text{llm}}(y | x)}{\pi_{\text{ref}}(y | x)} + \beta \log Z, \quad (10)$$

where $Z = \sum_{y'} \pi_{\text{ref}}(y' | x) \exp(\frac{1}{\beta} r(x, y'))$.

Then,

$$\begin{aligned}
 \Pr(y_w \geq y_l^{(1)}, \dots, y_w \geq y_l^{(m)}) &= \text{softmax}(r(x, y_w)) \\
 &= \frac{\exp(r(x, y_w))}{\exp(r(x, y_w)) + \sum_{i=1}^m \exp(r(x, y_l^{(i)}))} \\
 &= \frac{1}{1 + \sum_{i=1}^m \exp(r(x, y_l^{(i)}) - r(x, y_w))} \\
 &= \frac{1}{1 + \sum_{i=1}^m \exp(\gamma(y_l^{(i)} | x) + \beta \log Z - \gamma(y_w | x) - \beta \log Z)} \\
 &= \frac{1}{1 + \sum_{i=1}^m \exp(\gamma(y_l^{(i)} | x) - \gamma(y_w | x))} \\
 &= \frac{\exp(\gamma(y_w | x))}{\exp(\gamma(y_w | x)) + \sum_{i=1}^m \exp(\gamma(y_l^{(i)} | x))}
 \end{aligned} \tag{11}$$

We can learn π_θ by maximizing the logarithm-likelihood:

$$\max \log \Pr(y_w \geq y_l^{(1)}, \dots, y_w \geq y_l^{(m)}) \Leftrightarrow \min -\log \Pr(y_w \geq y_l^{(1)}, \dots, y_w \geq y_l^{(m)}) = \mathcal{L}, \tag{12}$$

$$\therefore \mathcal{L}_{\text{con}} = -\mathbb{E} \left[\log \frac{\exp(\gamma(y_w | x))}{\exp(\gamma(y_w | x)) + \sum_{i=1}^m \exp(\gamma(y_l^{(i)} | x))} \right], \tag{13}$$

$$\text{where } \gamma(y | x) = \beta \log \frac{\pi_\theta(y | x)}{\pi_{\text{ref}}(y | x)}. \tag{14}$$

F.2. LambdaRank ranking

Theorem F.2. Let x be a prompt and (y_1, \dots, y_m) be the responses for x under the LambdaRank assumption (Eq.(6)). Then the objective function to learn the LLM π_θ :

$$\mathcal{L}_{\text{lamb}} = -\mathbb{E} \left[\sum_{1 \leq i < j \leq m} \log \sigma(\gamma(y_i | x) - \gamma(y_j | x)) \right]. \tag{15}$$

Proof.

$$\begin{aligned}
 \Pr(y_1 \geq \dots \geq y_m) &= \prod_{1 \leq i < j \leq m} \sigma(r(x, y_i) - r(x, y_j)) \\
 &= \prod_{1 \leq i < j \leq m} \sigma(\gamma(x, y_i) + \beta \log Z - \gamma(x, y_j) - \beta \log Z) \\
 &= \prod_{1 \leq i < j \leq m} \sigma(\gamma(y_i | x) - \gamma(y_j | x)).
 \end{aligned} \tag{16}$$

We can learn π_θ by maximizing the logarithm-likelihood:

$$\max \log \Pr(y_w \geq y_l^{(1)}, \dots, y_w \geq y_l^{(m)}) \Leftrightarrow \min -\log \Pr(y_w \geq y_l^{(1)}, \dots, y_w \geq y_l^{(m)}) = \mathcal{L}, \tag{17}$$

$$\therefore \mathcal{L}_{\text{lamb}} = -\mathbb{E} \left[\sum_{1 \leq i < j \leq m} \log \sigma(\gamma(y_i | x) - \gamma(y_j | x)) \right], \tag{18}$$

$$\text{where } \gamma(y | x) = \beta \log \frac{\pi_\theta(y | x)}{\pi_{\text{ref}}(y | x)}. \tag{19}$$

F.3. ListMLE ranking

Theorem F.3. Let x be a prompt and (y_1, \dots, y_m) be the responses for x under the ListMLE assumption (Eq.(7)). Then the objective function to learn the LLM π_θ :

$$\mathcal{L}_{\text{lmle}} = -\mathbb{E} \left[\sum_{i=1}^m \log \frac{\exp(\gamma(y_i | x))}{\exp(\gamma(y_i | x)) + \sum_{j=i}^m \exp(\gamma(y_j | x))} \right]. \quad (20)$$

Proof. From Eq.(11),

$$\begin{aligned} \Pr(y_1 \geq \dots \geq y_m) &= \prod_{i=1}^m \Pr(y_i \geq y_{i+1}, \dots, y_i \geq y_m) \\ &= \prod_{i=1}^m \frac{\exp(\gamma(y_i | x))}{\exp(\gamma(y_i | x)) + \sum_{j=i+1}^m \exp(\gamma(y_j | x))}. \end{aligned} \quad (21)$$

We can learn π_θ by maximizing the logarithm-likelihood:

$$\max \log \Pr(y_w \geq y_l^{(1)}, \dots, y_w \geq y_l^{(m)}) \Leftrightarrow \min -\log \Pr(y_w \geq y_l^{(1)}, \dots, y_w \geq y_l^{(m)}) = \mathcal{L}, \quad (22)$$

$$\therefore \mathcal{L}_{\text{lmle}} = -\mathbb{E} \left[\sum_{i=1}^m \log \frac{\exp(\gamma(y_i | x))}{\exp(\gamma(y_i | x)) + \sum_{j=i+1}^m \exp(\gamma(y_j | x))} \right], \quad (23)$$

$$\text{where } \gamma(y | x) = \beta \log \frac{\pi_\theta(y | x)}{\pi_{\text{ref}}(y | x)}. \quad (24)$$

G. Baselines

We conduct detailed illustrations on the baselines compared with LARPO in Section 5 below.

- RRHF (Yuan et al., 2023) scores responses via a logarithm of conditional probabilities and learns to align these probabilities with human preferences through ranking loss.
- SLiC-HF (Zhao et al., 2023b) proposes a sequence likelihood calibration method which can learn from human preference data.
- DPO (Guo et al., 2024) simplifies the PPO (Ouyang et al., 2022) algorithms into an offline direct optimization objective with the pairwise Bradley-Terry assumption.
- IPO (Azar et al., 2024) theoretically grounds pairwise assumption in DPO into a pointwise reward.
- CPO (Xu et al., 2024a) adds a reward objective with sequence likelihood along with the SFT objective.
- KTO (Ethayarajh et al., 2024) adopts the Kahneman-Tversky model and proposes a method which directly maximizes the utility of generation instead of the likelihood of the preferences.
- RDPO (Park et al., 2024) modifies DPO by including an additional regularization term to disentangle the influence of length.
- SimPO (Meng et al., 2024b) further simplifies the DPO objective by using the average log probability of a sequence as the implicit reward and adding a target reward margin to the Bradley-Terry objective.
- Iterative DPO (Xiong et al., 2024) identifies the challenge of offline preference optimization and proposes an iterative learning framework.

H. Experiment settings

H.1. Table 2

We conduct evaluation on two widely used benchmark: AlpacaEval2 (Dubois et al., 2024) and MixEval (Ni et al., 2024). We consider two base models: Mistral-7b-base and Mistral-7b-it. For Mistral-7b-base, we first conduct supervised finetuning following Meng et al. (2024b) before the preference optimization.

Table 7. Preference optimization objective study on AlpacaEval2 and MixEval. For AlpacaEval2, we report the result with both opensource LLM evaluator `alpaca_eval_llama3-70b_fn` and GPT4 evaluator `alpaca_eval_gpt4_turbo_fn`. SFT corresponds to the initial chat model.

	Method	AlpacaEval 2 (opensource LLM)		AlpacaEval 2 (GPT-4)		MixEval	MixEval-Hard
		LC Winrate	Winrate	LC Winrate	Winrate	Score	Score
Gemma2-2b-it	SFT	47.03	48.38	36.39	38.26	0.6545	0.2980
	pairwise	55.06	66.56	41.39	54.60	0.6740	0.3375
	contrastive	60.44	72.35	43.41	56.83	0.6745	0.3315
	ListMLE	63.05	76.09	49.77	62.05	0.6715	0.3560
	LambdaRank	58.73	74.09	43.76	60.56	0.6750	0.3560
Mistral-7b-it	SFT	27.04	17.41	21.14	14.22	0.7070	0.3610
	pairwise	49.75	55.07	36.43	41.86	0.7175	0.4105
	contrastive	52.03	60.15	38.44	42.61	0.7260	0.4340
	ListMLE	48.84	56.73	38.02	43.03	0.7360	0.4200
	LambdaRank	51.98	59.73	40.29	46.21	0.7370	0.4400

The performance scores for offline preference optimization baselines are from SimPO (Meng et al., 2024b). To have a fair comparison with these baselines, we adopt the same off-the-shelf reward model (Jiang et al., 2023b) as in SimPO for the iterative DPO baseline and LARPO.

For the iterative DPO baseline, we generate 2 responses for each prompt, score them with the off-the-shelf reward model and construct the preference pair data to tune the model.

For LARPO (contrastive \mathcal{L}_{con}), we generate 10 responses each iteration and score them with the reward model. The top-1 ranked response and the bottom-3 ranked responses are adopted as the chose response and rejected responses respectively. Generation temperature is selected as 1 and 0.8 for Mistral-7b-base and Mistral-7b-it respectively (we search it among 0.8, 0.9, 1.0, 1.1, 1.2).

For LARPO (LambdaRank $\mathcal{L}_{\text{lamb}}$), we generate 10 responses each iteration and score them with the reward model. The top-2 ranked response and the bottom-2 ranked responses are adopted as the chose response and rejected responses respectively. Generation temperature is selected as 1 and 0.8 for Mistral-7b-base and Mistral-7b-it respectively (we search it among 0.8, 0.9, 1.0, 1.1, 1.2).

For LARPO (ListMLE $\mathcal{L}_{\text{lmle}}$), we generate 10 responses each iteration and score them with the reward model. The top-2 ranked response and the bottom-2 ranked responses are adopted as the chose response and rejected responses respectively. Generation temperature is selected as 1 and 0.8 for Mistral-7b-base and Mistral-7b-it respectively (we search it among 0.8, 0.9, 1.0, 1.1, 1.2).

LARPO can achieve even stronger performance with stronger off-the-shelf reward model (Dong et al., 2024).

H.2. Table 3

We conduct experiments on both Gemma2-2b-it (Team et al., 2024b) and Mistral-7b-it (Jiang et al., 2023a). Following Tunstall et al. and Dong et al. (2024), we perform training on UltraFeedback dataset for 3 iterations and show the performance of the final model checkpoint. We use the pretrained reward model from Dong et al. (2024). The learning rate is set as $5e-7$ and we train the LLM for 2 epochs per iteration.

For the pairwise objective, we generate 2 responses for each prompt and construct the preference pair data with the reward model. For the others, we generate 4 responses per prompt and rank them with the reward model. For the contrastive objective, we construct the 1-vs-N data with the top-1 ranked response and the other responses. For the listMLE and lambdarank objective, we take the top-2 as positives and the last-2 as the negatives. Experiments with opensource LLM as the evaluator (`alpaca_eval_llama3-70b_fn`) can be found in Table 7.

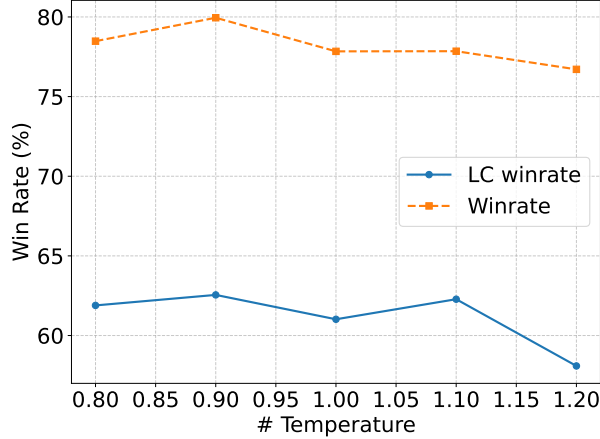


Figure 6. Training temperature study with $\mathcal{L}_{\text{pair}}$ on Gemma2-2b-it and Alpaca Eval 2. Within a specific range (> 0.9), lower temperature leads to harder negative and benefit the trained LLM. However, temperature lower than this range can cause preferred and rejected responses non-distinguishable and lead to degrade training.

H.3. Table 4

We adopt Gemma2-2b-it as the initial model. All the models are trained with iterative DPO for 3 iterations. We use the off-the-shelf reward model (Dong et al., 2024). We generate 2 responses for each prompt in each iteration. For “w. current”, we only use the scored responses in the current iteration for preference optimization data construction. For “w. current + prev”, we rank the responses in the current iteration and the previous one iteration, and construct the preference pair data with the top-1 and bottom-1 ranked responses. For “w. current + all prev”, we rank all the responses for the prompt in the current and previous iterations and construct the preference pair data. For “single temperature”, we only adopt temperature 1 and generate 2 responses for reward model scoring. For “diverse temperature”, we generate 2 responses with temperature 1 and 0.5 respective and rank the 4 responses to construct the preference data with the reward model.

H.4. Table 6

We use mistral-7b-it (Jiang et al., 2023a) as the initial model to alleviate the influence of the math related post-training data of the original model. For SFT, we conduct training on the meta-math dataset (Yu et al., 2023). For DPO, we use the prompts in the training set of the two dataset and conduct online iterative preference optimization with the binary rule-based reward (measure if the final answer is correct or not with string match). The evaluation is performed on the test set of MATH and GSM8K respectively. For SFT, we follow the same training setting with Yu et al. (2023). For DPO, we search the learning rate in $1e-7$, $2e-7$, $5e-7$, $2e-8$, $5e-8$ and train the LLM for 5 iterations with early stop (1 epoch per iteration for MATH and 2 epoch per iteration for GSM8K). The learning rate is set as $1e-7$ and we select the checkpoint after the first and fourth iteration for GSM8K and MATH respectively.

H.5. Figure 4(a)

We conduct training with the prompts in the training set of GSM8K and perform evaluation on GSM8K testing set. We conduct learning rate search and finalize it to be $2e-7$. The learning is performed for 3 iterations.

We make explanations of how we construct the four types of negative settings: For (1) a random response not related to the given prompt, we select a response for a random prompt in Ultrafeedback. For (2) a response to a related prompt, we pick up a response for a different prompt in the GSM8K training set. For (3) an incorrect response to the given prompt with high temperature, we select the temperature to be 1. For (4) an incorrect response to the given prompt with low temperature, we select the temperature to be 0.7.

H.6. Figure 4(b)

We conduct experiments on both Gemma2-2b-it and Mistral-7B-it models. For both LLMs, we conduct iterative DPO for 3 iterations and report the performance of the final model. We perform evaluation on Alpaca Eval2 with

`alpaca_eval_llama3_70b_fn` as the evaluator.

For temperature study, we find that under a specific temperature threshold, repeatedly generated responses will be large identical for all LLMs and cannot be used to construct preference data, while the threshold varies for different LLMs. The “low” and “high” refer to the value of those selected temperatures. We also conduct experiments on Gemma2-2b-it model and show the results in Figure 6.

H.7. Figure 4(c)

We adopt Mistral-7b-it as the initial LLM and the contrastive objective (Eq. 9) in iterative preference optimization. We generate 4/6/8/10 responses with the LLM and score the responses with the off-the-shelf reward model (Dong et al., 2024). The top-1 scored response is adopted as the positive response and the other responses are treated as the negative responses to construct the 1-vs-N training data. The temperature is set as 1 to generate the responses.