ICDPO: Effectively Borrowing Alignment Capability of Others via In-context Direct Preference Optimization

Feifan Song[1,2], **Yuxuan Fan**[1,2], **Xin Zhang**[3], **Peiyi Wang**[1,2], **Houfeng Wang**[1,2*]
[1]National Key Laboratory of Multimedia Information Processing, Peking University
[2]School of Computer Science, Peking University [3]Microsoft Research Asia
{songff,yxfan}@stu.pku.edu.cn; xinzhang3@microsoft.com
wangpeiyi9979@gmail.com; wanghf@pku.edu.cn

Abstract

Large Language Models (LLMs) rely on Human Preference Alignment (HPA) to ensure the generation of safe content. Due to the heavy cost associated with fine-tuning, fine-tuning-free methods have emerged, typically modifying LLM decoding with external auxiliary methods. However, these methods do not essentially enhance the LLM itself. In this paper, we rethink the derivation procedures of DPO, based on which we conversely build an instant scorer using the states of the LLM before and after In-context Learning (ICL). Accordingly, we propose a novel approach called In-Context Direct Preference Optimization (ICDPO). It enables LLMs to borrow the HPA capabilities from superior LLMs with ICL, generating well-aligned responses as estimated by the aforementioned instant scorer, thereby enhancing the final performance. ICDPO can be further enhanced with a two-stage retriever and an upgraded scorer, both offering benefits. Extensive experiments show its effectiveness, particularly in outperforming two fine-tuning-free baselines, and it exhibits competitiveness with SFT + LoRA. We also conduct detailed analyses to offer comprehensive insights into ICDPO.[1]

1 Introduction

Human Preference Alignment (HPA) is crucial within the LLM industry as it prevents LLMs from generating offensive, harmful, or misleading content contrary to human values. Presently, mainstream approaches to HPA heavily depend on fine-tuning, exemplified by RLHF (Stiennon et al., 2020; Ouyang et al., 2022; Zhu et al., 2023), RAFT (Dong et al., 2023a), RRHF (Yuan et al., 2023), or DPO (Rafailov et al., 2023). Nevertheless, the huge computational and data annotation costs associated with fine-tuning are hard to ignore.

As a response, fine-tuning-free approaches have gained popularity. Li et al. (2024) enable the LLM to take self-evaluation in decoding process. Alternatively, LLMs can borrow the capabilities of superior models (i.e. teacher models) to improve responses. Here the concept of *borrowing* is different from *learning* for it does not bring real parameter updates. For instance, external scorers capable of distinguishing human preference can be involved to apply best-of-N selection for multiple candidates or enhance block selection during LLM inference (Mudgal et al., 2023).

However, these approaches concentrate on the decoding stage, neglecting to fundamentally enhance the HPA capabilities of the LLM itself. This limitation raises the question: **Can LLMs borrow the HPA capabilities of superior LLMs to develop themselves without fine-tuning?** Therefore, we select In-context Learning (ICL) to reach the target of *borrowing*, as depicted in Figure 1(a). Unlike *learning*, ICL enables LLMs to ingest well-aligned samples from external teachers, mimicking them to produce aligned responses without fine-tuning.

More importantly, we rethink the procedures of Direct Preference Optimization (DPO) proposed in Rafailov et al. (2023). It integrates the policy LLM into the Reward Modeling by transforming RLHF objectives, bridging the relation between the provided reward model (RM) and optimal policy π^*. Here, the RM quantifies the distributional disparity between π^* and its reference model π_0. Conversely, an optimized policy that aligns with human preference can collaborate with its pre-optimized reference model, potentially offering more reliable estimations of HPA for candidate responses.

Additionally, LLMs essentially undergo instantaneous meta-optimization via ICL, involving an internal parameter updating formulation similar to real fine-tuning (Dai et al., 2023a). Consequently, the states of an LLM before and after ICL can be regarded as the **Expert** π^* and **Amateur** π_0, respec-

[*] Corresponding author.

[1] The code of this work is available at https://github.com/F2-Song/ICDPO.

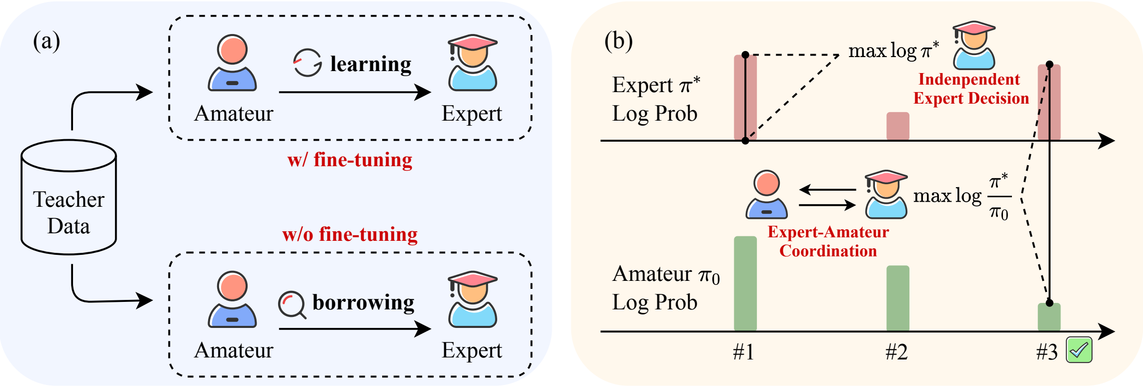

Figure 1: The overview of ICDPO. (a) The difference in teacher data utilization between normal fine-tuning and ICL without fine-tuning. (b) The core of ICDPO is that expert-amateur coordination maximizes S which represents the disparity between the expert and the amateur. It brings more accurate estimation than using only the expert LLM.

tively, to form a customized RM for scoring multiple samples (named Contrastive Score S), thereby maximizing the effectiveness of ICL, as illustrated in Figure 1(b). This process remains fine-tuning-free and entails only one LLM during decoding, which we term as **In**-**C**ontext **D**irect **P**references **O**ptimization (**ICDPO**).

Since we intend to harness the LLM through contextual demonstrations, the selection and ordering of demonstrated samples become crucial. Inspired by the nature of fine-tuning, where aligned distributions between training and test sets maximize effectiveness, we develop a two-stage retriever to identify demonstrations that are most similar to the test samples in both form and semantics, thereby improving the performance of ICDPO. Furthermore, like the prevalent contrastive fine-tuning in HPA, we elevate S to $\hat{S}$ by incorporating both favorable and unfavorable samples to amplify the disparities between π^* and π_0. It works as debiasing the distribution of candidates to further enhance ICDPO.

Extensive experiments are conducted to evaluate the proposed ICDPO, encompassing evaluations using both RM and GPT-4, along with an ablation study to validate each module. We also provide comprehensive analyses of multiple aspects in ICDPO. The main observations are as follows:
(1) ICDPO borrows the HPA ability from superior LLMs through ICL, which in turn produces the π^* collaborating with the initial π_0 to conduct scoring. This significantly enhances performance by improving and exploiting the LLM itself, surpassing two fine-tuning-free baselines, as well as being competitive with SFT plus LoRA(Hu et al., 2022).
(2) Contextual demonstrations are closely related to the final performance. Specifically, demonstrated samples of higher quality and the proposed two-stage retriever can both facilitate ICDPO.
(3) Regarding scoring, the scorers S and $\hat{S}$ in ICDPO can provide reliable estimations of the degree of HPA, which can also be applied to fine-tuning methods, like DPO. *How to measure HPA (human preference alignment?)*

2 Methodology

In this section, we rethink the transformation from RLHF to DPO (Rafailov et al., 2023), an elegant supervised fine-tuning algorithm derived from the original RLHF objective $\mathcal{T}$. We focus on the relation between a given RM and the corresponding optimal policy π^*, and adapt it to LLM inference in the manner of In-context Learning (ICL), which we term as ICDPO.

2.1 From Reward Model to Policy LLM

The original target $\mathcal{T}$ of RLHF is to optimize the policy LLM π for the acquisition of a synthetic reward $\mathcal{R}$, the combination of a fundamental reward from the given RM r^* and a KL-regularization to reference policy π_0,

$$\begin{aligned}\mathcal{T} &= \max_{\pi} \mathbb{E}[\mathcal{R}] \\ &= \max_{\pi} \mathbb{E}[r^*(x,y) - \beta \log \frac{\pi(y \mid x)}{\pi_0(y \mid x)}]\end{aligned} \tag{1}$$

Rafailov et al. (2023) construct the Direct Preference Optimization (DPO) algorithm by first transforming Equation 1,

$$\begin{aligned}\mathcal{T} &= \min_{\pi} \mathbb{E}[\log \frac{\pi(y \mid x)}{\pi_0(y \mid x)} - \frac{1}{\beta} r^*(x,y)] \\ &= \min_{\pi} \mathbb{E}[\log \frac{\pi(y \mid x) Z(x)}{\pi_0(y \mid x) \exp\left(\frac{1}{\beta} r^*(x,y)\right)} \\ &\quad - \log Z(x)]\end{aligned} \tag{2}$$

where

$$Z(x) = \sum_y \pi_0(y \mid x) \exp\left(\frac{1}{\beta} r^*(x, y)\right) \quad (3)$$

is the partition function, and the relation between r^* and the optimal policy π^* of Equation 2 is found:

$$r^*(x, y) = \beta \log \frac{\pi^*(y \mid x)}{\pi_0(y \mid x)} + \beta \log Z(x) \quad (4)$$

2.2 Preference Optimization via ICL

In RLHF, r^* typically represents the outcome of Reward Modeling preceding the PPO stage, and π^* denotes the corresponding optimal policy. DPO opts to integrate π into the supervised objective of Reward Modeling and devises an SFT-style fine-tuning approach based on the formulation of Equation 4. Conversely, we rethink Equation 1 and 4 with the aim of avoiding parameter modification in the policy LLM π.

With an optimized policy LLM π^* and a reference policy π_0, according to Equation 4, we can build a customized reward function $\hat{r}$ as follows:

$$\hat{r}(x, y) = \log \frac{\pi^*(y \mid x)}{\pi_0(y \mid x)} + \log Z(x) \quad (5)$$

Since π^* has been optimized to align with human preference, the corresponding $\hat{r}$ should well reflect the extent of human preference to some degree. Additionally, the synthetic $\mathcal{R}$ in Equation 1 incorporates the KL-regularization component to prevent the policy from deviating too far from the typical linguistic space. Therefore, if π^* is presumed to retain this capability without the concern for regularization, Equation 1 could exclusively concentrate on preference rewards. Consequently, with Equation 5, we could have

$$\begin{aligned} \max_y \mathcal{R} &\equiv \max_y \hat{r}(x, y) \\ &\equiv \max_y \log \frac{\pi^*(y \mid x)}{\pi_0(y \mid x)} \end{aligned} \quad (6)$$

because $Z(x)$ in Equation 5 involves only x.

Furthermore, π^* ought to be optimized while the initial objective necessitates it not to be fine-tuned. We thus use ICL to fulfill all these criteria, with inspiration from Dai et al. (2023a) that inner meta-optimization can be demonstrated in ICL with contextual demonstrations $\mathbf{d}$ and tested x:

$$\begin{aligned} &\text{Attention}([\mathbf{d}; x], q) \\ &\approx W_V [\mathbf{d}; x] (W_K [\mathbf{d}; x])^T q \\ &= \left(W_V x (W_K x)^T + W_V \mathbf{d} (W_K \mathbf{d})^T\right) q \\ &= (W_{\text{ZSL}} + \Delta W_{\text{ICL}})\, q \end{aligned} \quad (7)$$

Here, $q = W_Q t$ represents the query of the next token t in the self-attention mechanism, and $W_{\text{ZSL}} q = W_V x (W_K x)^T q$ approximates the attention result in a zero-shot setting (i.e., no demonstrations involved). Furthermore, $\Delta W_{\text{ICL}} = W_V \mathbf{d} (W_K \mathbf{d})^T$ updates the weights of W_{ZSL} using demonstrations $\mathbf{d}$ in the context, thereby facilitating meta-optimization.

As a result, the optimized π^* can be built directly through ICL, while the reference LLM π_0 serves as the initial checkpoint, i.e., the base model in this scenario. Moreover, π^* does not undergo parameter updates from fine-tuning, thereby preserving the initial language modeling capacity as π_0, without the need for additional regularization. Therefore, we can employ a two-stage inference pipeline. In the first stage, multiple responses $\mathbf{y}$ are sampled from π^* as candidates to guarantee a potentially acceptable output, termed as **Generation**. Subsequently, in the second **Scoring** stage, the contrastive score S for each candidate $y \in \mathbf{y}$ is computed based on the demonstrated samples $\mathbf{d}$, the prompt x, and Equation 6:

$$\begin{aligned} S(\mathbf{d}, x, y) &= \log \frac{\pi^*(y \mid x)}{\pi_0(y \mid x)} \\ &= \log \frac{\pi(y \mid [\mathbf{d}; x])}{\pi(y \mid x)} \end{aligned} \quad (8)$$

wherein the most preferred response y^* can be chosen based on the largest S, indicating the highest reward of human preference, as in Figure 1(b). We summarize the entire workflow as **ICDPO**. Note that π^* is acquired through ICL, implying that only a single checkpoint is required throughout the entire inference process. We define the score of response y towards prompt x from π as its probability of generating y,

$$\pi(y \mid x) = \sum_i P_\pi(y_i | x, y_{<i}) \quad (9)$$

2.3 Connection to Contrastive Decoding

We observe that Equation 6 relies on a contrastive estimation involving two LLMs: π^* and π_0. Furthermore, Li et al. (2023) enhance the quality of generated texts by replacing the naive maximum probability decoding with a contrastive objective, namely Contrastive Decoding (CD), where each step utilizes both an expert model π^+ and an amateur model π^-,

$$y_i^* = \arg\max_{y_i} \log \frac{\pi^+(y_i \mid x, y_{<i})}{\pi^-(y_i \mid x, y_{<i})} \quad (10)$$

Algorithm 1: ICDPO

```
Input: Language Model $\pi$, Dataset $D$,
       input prompt $x$   such query has annotated output in the dataset D
Output: Response $y$ with the largest score
// Generation stage
Retrieve $m$ demonstrated samples $\mathbf{d}$ from $D$
Sample $n$ responses $\{y_i\}$ from $\pi(y \mid [\mathbf{d}; x])$
// Scoring stage
Let $s = -\infty$
Let $p = 0$
for $y_i \in \{y_1, ..., y_n\}$ do
    Estimate $\pi(y \mid [\mathbf{d}; x])$ in ICL
    Estimate $\pi(y \mid x)$
    Estimate $S(\mathbf{d}, x, y)$ with Equation 8
    if $S(\boldsymbol{d}, x, y) > s$ then
        $s = S(\mathbf{d}, x, y)$
        $p = i$
    end if
end for
Let $y = y_p$
return $y$
```

While Equation 6 optimizes at the sentence-level instead of estimating token-wise scores as in CD for the generated y, we note that π^* and π_0 are essentially treated as the expert and amateur models, respectively, in terms of HPA. This enhances LLM decoding with a focus on human preference. To achieve this, we can enhance Equation 6 and Equation 8 by introducing a purposely worse policy π^- for HPA to replace the original π_0. More precisely, π^- can also be acquired through In-context Learning with human-rejected samples $\mathbf{d}^-$ as demonstrations, whereas the original expert model π^* in Equation 6 can be relabeled as π^+ and its contextual demonstrations comprise solely human-chosen $\mathbf{d}^+$. Hence, the promoted contrastive score is

$$\begin{aligned}\hat{S}(\mathbf{d}^+, \mathbf{d}^-, x, y) &= \log \frac{\pi^+(y \mid x)}{\pi^-(y \mid x)} \\ &= \log \frac{\pi(y \mid [\mathbf{d}^+; x])}{\pi(y \mid [\mathbf{d}^-; x])}\end{aligned} \tag{11}$$

2.4 Retrieval

The demonstrated samples and their sequencing are acknowledged as crucial factors for ICL. Since the process of ICL may resemble gradient descent during actual model training, we can further amplify the inner meta-optimization from the fine-tuning standpoint. Given that the closeness between the distributions of the test data and the training data is vital for the efficacy of fine-tuning, it should coherently work in ICL. Consequently, we also employ a prevalent similarity-based retriever to determine the sample selection and their corresponding sequencing, while incorporating additional considerations: (1) Despite their effectiveness, pre-trained retrievers (e.g., SBERT-based methods) have significant computational costs for the large number of samples, requiring a two-stage design where coarse-grained selections are first made before more fine-grained retrievals. (2) Since LLMs operate in an auto-regressive manner, the last portion of the tested samples should have the most significant impact. Hence, retrieving those with structurally similar end portions is prioritized, and able to additionally reduce computational overhead.

Method	**Harmless**	**Helpful**	**Total**
w/ *HH-RLHF*$_{\text{raw}}$			
Raw	24.23	-47.62	-11.70
LLaMA2-chat	105.97	61.18	83.57
GPT-3.5-turbo	105.99	73.80	89.89
w/ *SytheticGPT*$_{\text{raw}}$			
Raw	-	-	74.04
LLaMA2-chat	-	-	120.31

Table 1: Reference results.

Therefore, we propose a two-stage retriever containing a coarse-grained BM25 retriever (Robertson and Zaragoza, 2009) focusing on the end of each sample, and an SBERT (Reimers and Gurevych, 2019) to execute fine-grained retrieval:

$$\begin{aligned}R(\{x_i\}) &= \text{SBERT}(\{a_j\}) \\ \{a_j\} &= \text{BM25}(\{x_i[-L:]\})\end{aligned} \tag{12}$$

where $\{x_i\}$ is the support set, and L is the window size constraining the ending range of samples for BM25. We show that ICDPO equipped with $\mathcal{R}$ yields notable improvement overall.

3 Experiment

3.1 Settings

We employ two datasets, *HH-RLHF* and *SyntheticGPT* to comprehensively assess the effectiveness of ICDPO. Regarding the superior teacher models, we included LLaMA2-7B-chat (denoted as **LLaMA2-chat**) and **GPT-3.5-turbo** to support all methods with base models. For *HH-RLHF*, we present the original version (referred

Method	LLaMA			LLaMA2			Mistral		
	Harmless	Helpful	Total	Harmless	Helpful	Total	Harmless	Helpful	Total
w/ *HH-RLHF*$_{\text{raw}}$									
Base	4.47	-77.53	-36.54	6.25	-67.67	-30.72	9.59	-33.22	-11.82
SFT	20.23	-65.46	-22.63	20.48	-60.77	-20.16	21.91	-48.65	-13.38
ICDPO	25.02	-64.95	-19.97	39.81	-71.89	-16.05	26.60	-51.38	-12.40
ICDPO+$\hat{S}$	24.03	-55.86	-15.92	42.52	-63.53	-10.52	32.78	-42.82	-5.03
ICDPO+$\hat{S}R$	22.50	-55.77	-16.64	31.54	-63.22	-15.85	25.15	-44.75	-9.81
w/ LLaMA2-chat									
SFT	48.92	20.54	34.73	72.94	42.24	57.59	77.59	49.29	63.43
RM-Aug	5.06	-60.35	-27.66	2.92	-52.12	-24.61	13.65	-7.00	3.32
RM-BoN	-1.47	-60.60	-31.04	2.90	-48.53	-22.82	7.16	-6.11	0.52
ICDPO	68.75	-17.61	25.56	97.06	27.49	62.27	99.29	38.34	68.81
ICDPO+$\hat{S}$	68.73	-11.75	28.48	98.03	29.36	63.69	97.26	45.08	71.16
ICDPO+$\hat{S}R$	90.54	12.59	51.56	101.08	38.26	69.66	101.68	45.51	73.59
w/ GPT-3.5-turbo									
SFT	54.28	-16.17	19.05	72.72	33.03	52.87	90.98	62.00	76.49
ICDPO	63.91	-23.27	20.31	91.56	16.33	53.94	85.10	21.23	53.16
ICDPO+$\hat{S}$	64.03	-14.86	24.58	92.14	21.40	56.76	85.83	36.14	60.98
ICDPO+$\hat{S}R$	82.21	3.63	42.91	98.77	28.08	63.42	92.21	39.55	65.88

Table 2: Main results on *HH-RLHF* scored by RM$_{\text{test}}$. **Higher** values represent **better** performance towards HPA.

to as *HH-RLHF*raw) and its enhanced version from LLaMA2-chat and GPT-3.5-turbo, while for *SyntheticGPT*, we consider both the original version (referred to as *SyntheticGPT*raw) and the version adapted from LLaMA2-chat.

We implement three base models for comprehensive evaluation: LLaMA-7B (Touvron et al., 2023a), LLaMA-2-7B (Touvron et al., 2023b), and Mistral-7B-v0.1 (Jiang et al., 2023), which we label as **LLaMA**, **LLaMA2**, and **Mistral**, respectively. The details of data preparation and implementation (including the reward model RM$_{\text{test}}$) can be found in Appendix A and B, respectively. We evaluate the performance of both the original candidates and new ones from teacher models in these datasets using RM$_{\text{test}}$, as presented in Table 1.

3.2 Main Results

Automatic evaluations are conducted on both *HH-RLHF* and *SyntheticGPT*. We deploy base models and their SFT variants on each dataset, utilizing LoRA (Hu et al., 2022) to accommodate the limitations of constrained devices. Since ICDPO essentially borrows the capabilities of superior LLMs, we also deploy two *borrowing* baselines, RM-BoN and RM-Aug, based on the Best-of-N policy and Mudgal et al. (2023), respectively. RM-BoN and RM-Aug can utilize the logits of superior LLMs as the external scorer (Fu et al., 2023) to select the best response or intermediate block during decoding. Although we introduce both LLaMA2-chat and GPT-3.5-turbo as the teachers, the detailed log probability of prompt tokens from GPT-3.5-turbo appears to be **inaccessible**, so we must compare ICDPO and the two baselines using only LLaMA2-chat on *HH-RLHF* and *SyntheticGPT*.

As to ICDPO, we evaluate its original version (supported by randomly sampled demonstrations) and variants with only $\hat{S}$ or both $\hat{S}$ and retriever R. We accordingly set the following research questions (RQs) to guide experiments:

RQ1: How does ICDPO perform well?

Table 2 presents the main results for *HH-RLHF*, while those for *SyntheticGPT* are provided in Appendix C. Essentially, all methods show notable improvements over the corresponding base models. However, in the specific scenario where LLaMA2-chat is referenced, ICDPO exhibits significant progress compared to RM-Aug and RM-BoN. Overall, ICDPO generally demonstrates competitive performance against SFT despite not undergoing fine-tuning. These results strongly support the effectiveness of ICDPO.

Method	LLaMA			LLaMA2			Mistral		
	Harmless	Helpful	Total	Harmless	Helpful	Total	Harmless	Helpful	Total
				w/ LLaMA2-chat					
ICDPO+R	90.55	9.96	50.24	100.62	35.89	68.25	101.49	40.34	70.91
ICDPO+BM25	84.99	3.18	44.08	99.78	31.89	65.83	102.54	43.74	73.13
ICDPO	68.75	-17.61	25.56	97.06	27.49	62.27	99.29	38.34	68.81
ICL	62.30	-26.09	18.09	97.23	16.72	56.97	94.79	32.68	63.73
ICL$_{\text{uni}}$	63.04	-25.25	18.89	95.64	14.74	55.18	94.54	33.06	63.80
				w/ GPT-3.5-turbo					
ICDPO+R	80.64	-1.13	39.75	98.08	24.45	61.25	89.91	31.10	60.50
ICDPO+BM25	74.28	-3.24	35.51	96.18	25.96	61.06	88.30	30.73	59.50
ICDPO	63.91	-23.27	20.31	91.56	16.33	53.94	85.10	21.23	53.16
ICL	52.73	-32.05	10.33	88.00	4.74	46.36	75.46	16.38	45.91
ICL$_{\text{uni}}$	50.85	-33.44	8.70	88.62	2.16	45.38	72.72	15.32	45.51

Table 3: Ablation study on *HH-RLHF*.

Furthermore, we observed that each method could receive lower scores in the domain of **Helpful** compared to **Harmless**. We infer that **Helpful** needs more substantial content from base models or external sources, whereas **Harmless** may only require simpler stylistic changes. Thus, Mistral, being the superior model combined with SFT where downstream information is forcibly integrated, achieves the highest scores in the **Helpful** domain. However, ICDPO also effectively enhances **Helpful** for Mistral, activated by contextual demonstrations, which is second only to SFT.

RQ2: How demonstrations affect ICDPO?

Intuitively, the quality of data, i.e. HPA degree, should heavily impact performance. For instance, GPT-3.5-turbo can generally provide greater assistance for SFT with higher-quality samples compared to ordinary sources, as proved in Song et al. (2023). ICDPO hereby reflects similar trends. According to Table 1, GPT-3.5-turbo and/or LLaMA2-chat can achieve higher scores than the original samples, consistent with Table 2 where ICDPO demonstrates improvements from superior demonstrations. This suggests that the meta-optimization in ICL does indeed function. In § 3.3, we will provide a detailed analysis of the effects of S using these higher-quality demonstrations.

Despite GPT-3.5-turbo being more powerful than LLaMA2-chat according to Table 1, ICDPO seems better with demonstrations from LLaMA2-chat than GPT-3.5-turbo. Believing it is not a coincidence, we make further analyses in Appendix E.

RQ3: The impact of extra modules?

ICDPO relies on S and randomly sampled demonstrations by default. In Table 2, we also test ICDPO with only $\hat{S}$, or $\hat{S} + R$ which additionally involves the retriever R. The overall performance can be improved step by step, except that R with samples from the original datasets fails. We attribute these results to the quality of the samples, as R essentially narrows the gap between demonstrations and the tested sample. Thus, if the initially chosen/rejected samples are not sufficiently *good*/*bad*, the estimation of S collapses, and R further exacerbates the confusion through meta-optimization.

3.3 Ablation Study

In this section, we test the effectiveness of the remaining modules. Our experiments focus on the variants of *HH-RLHF* derived from LLaMA2-chat and GPT-3.5-turbo, as presented in Table 3.

Retriever R We analyze the impact of fine-grained and coarse-grained retrieval with SBERT and BM25, respectively. The results indicate that the latter approach (ICDPO+BM25 *vs.* ICDPO) can strongly enhance the meta-optimization in ICL, similar to genuine fine-tuning. However, the former one (ICDPO+R *vs.* ICDPO+BM25) occasionally results in marginal improvement (LLaMA2/Mistral on *HH-RLHF*+GPT-3.5-turbo) or even a decline (Mistral on *HH-RLHF*+LLaMA2-chat). They occur upon powerful LLMs (e.g. LLaMA2/Mistral against LLaMA) achieving high performance without SBERT, indicating that fine-grained retrieval provides greater benefits to weaker LLMs for strong LLMs can directly handle ICL well.

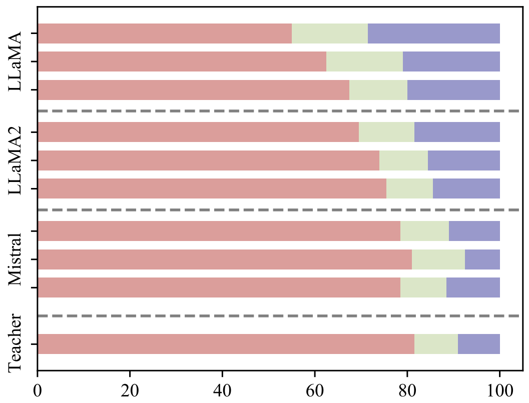

Figure 2: GPT-4 computed win-rates of ICDPO against golden responses in *HH-RLHF*, using demonstrations from the teacher (i.e. LLaMA2-chat). For each block titled by one base model, the bars from top to bottom are ICDPO, ICDPO+$\hat{S}$ and ICDPO+$\hat{S}R$, while **red**, **light green** and **purple** represent the proportion of **win**, **tie** and **lose**, respectively.

Contrastive Score S Without S, ICDPO degenerates into the normal ICL. We thus experiment with two decoding strategies: randomly selecting 1 from 3 candidates, and generating just 1 candidate[2]. Obviously, ICL without selections from S experiences significant performance declines, regardless of the decoding strategies. This validates the significance of S as the key element in ICDPO. Since S is a potential ranker, we also evaluate its performance in this aspect, as discussed in § 4.2.

3.4 GPT-4 Evaluation

We implement GPT-4 evaluation as an additional validation of automatic evaluation with RM$_{\text{test}}$, following Song et al. (2023); Liu et al. (2023b). We randomly select 200 samples from the test sets of *HH-RLHF* and evaluate ICDPO, ICDPO+$\hat{S}$, ICDPO+$\hat{S}R$, and their corresponding teachers. Their decoded responses are compared with the annotated choices in *HH-RLHF*$_{\text{raw}}$ to compute the win rate. In Figure 2, we use demonstrations from LLaMA2-chat for ICDPO, with LLaMA2-chat serving as the teacher model. The results for GPT-3.5-turbo can be found in Appendix D.

Initially, we consider placing the tested candidates in the prompt from double directions to mitigate positional bias, as discussed in Wang et al. (2023c). However, several attempts yield similar results regardless of the direction. We attribute it to the enhanced capabilities of GPT-4-32K and therefore use uni-directional tests to reduce costs.

Method	LLaMA	LLaMA2	Mistral
	w/ LLaMA2-chat		
SFT	34.73	57.59	63.43
DPO	43.02	68.34	69.26
DPO+S	48.11	71.62	71.84
	w/ GPT-3.5-turbo		
SFT	19.05	52.87	76.49
DPO	30.88	95.00	86.61
DPO+S	40.15	95.58	90.73

Table 4: DPO results on *HH-RLHF*.

We note that the results in Figure 2 align with those in Table 2, thereby validating the fairness of RM$_{\text{test}}$. Generally, ICDPO with $\hat{S}$ and R outperforms ICDPO without them. With the more powerful base model, the third block (Mistral) can even approach the performance of LLaMA2-chat.

4 Discussion

4.1 Extension of Contrastive Score

The contrastive score S utilizes the optimized π^* and initial π_0 to sort the candidates. Since ICL can be one of the implementation methods for π^*, other methods should also be able to utilize S.

Consequently, we implement DPO + LoRA using the TRL package (von Werra et al., 2020), with π defined as the n-th root of $P_\pi(y \mid x)$ to match the definition in DPO. We evaluate the performance with and without S (Table 4), demonstrating that S can still enhance DPO. It indicates that S may be a promising way for general use in HPA.

4.2 Consistency of Scoring

ICDPO computes the contrastive score S to rank sampled candidates y from ICL for the prompt x, similar to the methodology of RM$_{\text{test}}$. Therefore, we intend to evaluate ICDPO as the ranking model.

We introduce ICDPO, its enhanced version, ICDPO+$\hat{S}$, and its simplified variant (i.e., using only π^* for scoring, denoted as ICL), alongside RM$_{\text{test}}$. LLaMA2-chat is also incorporated as a reward model, like how it is used in RM-Aug and RM-BoN. We set up two scenarios: one depicted in Figure 3(a), where demonstrations for ICDPO are randomly selected, and the other depicted in Figure 3(b), which involves the proposed retriever R. In each scenario, we select 200 samples, each

[2] We also evaluate greedy search, which exhibits similar performance.

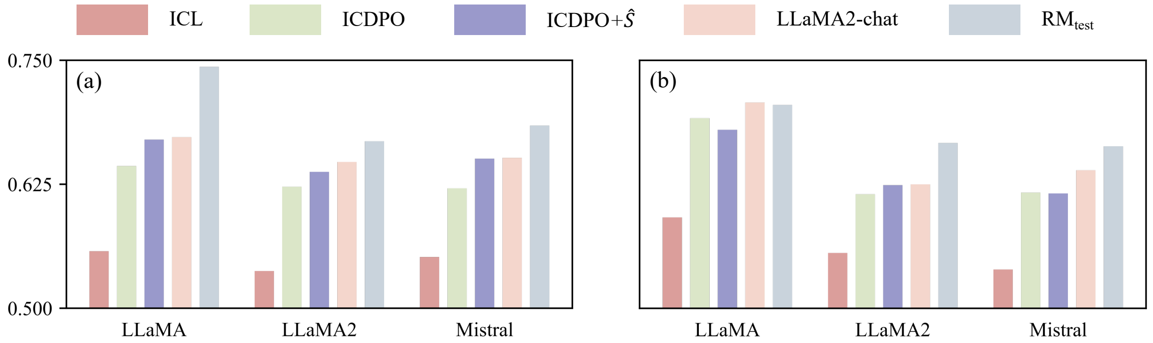

Figure 3: Results of consistency between different scorers and GPT-4. We compute MRR to measure the degree of consistency. (a) Results with randomly selected demonstrations. (b) Results with demonstrations retrieved by R.

containing 3 candidate responses sampled from the base model through ICL and sorted by GPT-4 as the ground truth. We use the Mean Reciprocal Rank (MRR) as the metric to fairly evaluate the competence of each method as a scorer and ranker.

Figure 3 illustrates that RM$_{\text{test}}$ achieves the highest performance in most cases, followed by LLaMA2-chat. ICDPO also performs well, with ICDPO+$\hat{S}$ generally yielding equal or higher MRR scores, even approaching the performance of LLaMA2-chat as the teacher. However, the performance of π^* itself is unsatisfactory, significantly lagging behind others. These findings exhibit that ICDPO is a potent scorer beyond the vanilla ICL and approaches the performance of LLaMA2-chat through effective *borrowing*.

5 Related Work

5.1 Human Preference Alignment

To mitigate the risk of generating toxic content, LLM should be aligned with human preference (Wang et al., 2023d), i.e. Human preference alignment (HPA), which has been advanced through RLHF (Ouyang et al., 2022; Zhu et al., 2024; Yu et al., 2023; Jang et al., 2023; Dai et al., 2023b) and SFT methods (Yuan et al., 2023; Song et al., 2023; Rafailov et al., 2023; Wang et al., 2023b; Zhang et al., 2023; Liu et al., 2023a; Xu et al., 2023; Hong et al., 2023). DPO (Rafailov et al., 2023) can be the representative one. It builds the relation between the RM and the combination of pre/post-optimized policies by transforming RLHF objective, which is inserted into reward modeling to derive an elegant SFT objective.

Nevertheless, fine-tuning LLMs is still costly. It triggers the need for fine-tuning-free methods, relying on self-selection (Li et al., 2024), external expert selection (Mudgal et al., 2023) or refinement of prompts (Cheng et al., 2023). The proposed ICDPO similarly refers to external experts, but does selection with self-estimation, which is based on reverse derivation of the relation in DPO.

5.2 In-Context Learning

LLM has the potential of instant few-shot learning through demonstrations in the context (Brown et al., 2020; **?**; Dong et al., 2023b; Zheng et al., 2023; Yang et al., 2023b), named In-Context Learning (ICL). The underlying mechanism of ICL has also been carefully studied. From the perspective of information flow, Wang et al. (2023a) distinguish the different roles of upper and lower layers in LLMs for ICL, while Dai et al. (2023a) established a dual relation between gradient descent and Transformer attention, thus illustrating that ICL as a meta-optimizer can be similar to explicit fine-tuning. We extend it to HPA, where the optimized policy can be easily acquired for generation and scoring without fine-tuning.

6 Conclusion

In this paper, we equip LLMs with HPA by leveraging capabilities from superior models without the need for costly fine-tuning. We rethink the procedures of DPO and focus on the crucial relation between the RM and the optimized policy. Building upon this relation, we propose ICDPO. It implements ICL to instantly optimize the LLM, which through collaboration with the initial policy can effectively estimate the degree of HPA and enhance the final performance. Comprehensive experiments demonstrate the effectiveness of ICDPO across various forms, encompassing both content generation and scoring. We hope this work to be a catalyst for further exploration of fine-tuning-free methods towards HPA.

7 Ethics Statement

We observe that the data involved in this work may indispensably contain sensitive, offensive, and misleading content, whose presence does not represent our attitudes, but is solely for research and should not be used or distributed outside of research contexts.

We are committed to establishing a more inclusive and ethically sound era of AI technology, which can be applied to legitimate needs and generate content that aligns with universally positive human values.

8 Limitations

ICDPO has been shown powerful but user-friendly, because it is fine-tuning-free and learns effectively from just demonstrations from superior LLMs. Although we conduct abundant experiments to evaluate ICDPO comprehensively, there remain a few aspects of limitation:
1. Despite 7B LLMs showing the satisfying capability of ICL, we fail to evaluate ICDPO on larger models for their costly requirements on hardware.
2. Similarly, we do not test the effect of changes in the number of demonstrations for ICL. Nonetheless, we believe it should further boost ICDPO with increasing demonstrations.
Due to limited computational resources, we leave them to the community with interest for further exploration.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jiale Cheng, Xiao Liu, Kehan Zheng, Pei Ke, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Black-box prompt optimization: Aligning large language models without model training. *arXiv preprint arXiv:2311.04155*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023a. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, Toronto, Canada. Association for Computational Linguistics.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2023b. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. 2023a. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2023b. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Jixiang Hong, Quan Tu, Changyu Chen, Xing Gao, Ji Zhang, and Rui Yan. 2023. Cyclealign: Iterative distillation from black-box llm to white-box models for better human alignment. *arXiv preprint arXiv:2310.16271*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Tiansheng Huang, Sihao Hu, and Ling Liu. 2024. Vaccine: Perturbation-aware alignment for large language model. *arXiv preprint arXiv:2402.01109*.

Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Yuhui Li, Fangyun Wei, Jinjing Zhao, Chao Zhang, and Hongyang Zhang. 2024. Rain: Your language models can align themselves without finetuning. In *International Conference on Learning Representations*.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. 2023a. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2023b. Aligning large language models with human preferences through representation engineering. *arXiv preprint arXiv:2312.15997*.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, et al. 2023. Controlled decoding from language models. *arXiv preprint arXiv:2310.17022*.

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, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2023. Preference ranking optimization for human alignment. *arXiv preprint arXiv:2306.17492*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, and Shengyi Huang. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023a. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, Singapore. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Tianyu Liu, and Zhifang Sui. 2023b. Making large language models better reasoners with alignment. *arXiv preprint arXiv:2309.02144*.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023c. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023d. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Weiwen Xu, Deng Cai, Zhisong Zhang, Wai Lam, and Shuming Shi. 2023. Reasons to reject? aligning language models with judgments. *arXiv preprint arXiv:2312.14591*.

Jiaxi Yang, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023a. Iterative forward tuning boosts in-context learning in language models. *arXiv preprint arXiv:2305.13016*.

Zhe Yang, Damai Dai, Peiyi Wang, and Zhifang Sui. 2023b. Not all demonstration examples are equally beneficial: Reweighting demonstration examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13209–13221, Singapore. Association for Computational Linguistics.

Tianshu Yu, Ting-En Lin, Yuchuan Wu, Min Yang, Fei Huang, and Yongbin Li. 2023. Constructive large language models alignment with diverse feedback. *arXiv preprint arXiv:2310.06450*.

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*.

Yichi Zhang, Zhuo Chen, Yin Fang, Lei Cheng, Yanxi Lu, Fangming Li, Wen Zhang, and Huajun Chen. 2023. Knowledgeable preference alignment for llms in domain-specific question answering. *arXiv preprint arXiv:2311.06503*.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4862–4876, Singapore. Association for Computational Linguistics.

Banghua Zhu, Jiantao Jiao, and Michael I Jordan. 2023. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*.

Banghua Zhu, Michael I Jordan, and Jiantao Jiao. 2024. Iterative data smoothing: Mitigating reward overfitting and overoptimization in rlhf. *arXiv preprint arXiv:2401.16335*.

A Dataset Preparation

We introduce the following two datasets for ICDPO:

- *HH-RLHF* is proposed by Bai et al. (2022), focusing on the domain of harmlessness and helpfulness in multi-turn conversations. While it initially consists of four subsets, we select two representative ones: *harmless-base* and *helpful-base*, which we denote as **Harmless** and **Helpful**, respectively. We mix the data of two domains for training, while separately evaluating each method in the main experiment.
- *SyntheticGPT*[3] collects about 33.1K samples of instruction following. Since its original version just has a training set, we manually split it into train/dev/test ones.

Each sample in these datasets has two candidates, including a shared prompt and two chosen/rejected candidate responses. In order to alleviate the pressure of GPU memory and accelerate the inference, we filter all samples according to sequence length in advance, 320/128 tokens for prompts/responses in *HH-RLHF*, while 128/200 in *SyntheticGPT*.

B Implementation Details

We implement ICDPO with all base models on Huggingface.Library (Wolf et al., 2020). For ICL, the number of demonstrations and top-p sampling is 2 and 3, respectively, where p is set to 0.8. To facilitate demonstration retrieval in ICL, we deploy BM25 and SBERT[4]. The BM25 model first retrieves 20 samples, which are then re-ranked by the SBERT retriever to obtain highly semantically similar ones. The templates for ICL have been placed in Appendix F for a detailed overview.

idea:
get various xxx level, annotated by human, then use semantically similarity to decide on the scoring of the sentiment

Furthermore, the third-party reward model for automatic scoring is denoted as RM_{test}[5], while GPT-4-32K is employed for GPT-4 evaluation. To carry out *borrowing* in ICL, we employ LLaMA2-chat to generate new choices for *HH-RLHF* and *SyntheticGPT*, while for GPT-3.5-turbo, we reuse $\textit{HH-RLHF}_{\text{ChatGPT,3}}$ released by Song et al. (2023). The whole details can be found in the released code.

C Additional Main Results

Method	**LLaMA**	**LLaMA2**	**Mistral**
	w/ *SytheticGPT*$_{\text{raw}}$		
Base	-121.85	-101.01	34.75
SFT	36.05	77.04	96.10
ICDPO	27.74	77.89	68.97
ICDPO+$\hat{S}$	29.37	83.34	78.34
ICDPO+$\hat{S}R$	53.97	72.26	71.69
	w/ LLaMA2-chat		
SFT	41.89	99.21	99.09
RM-Aug	-95.63	-77.14	84.48
RM-BoN	-97.38	-70.08	88.00
ICDPO	49.82	100.41	113.67
ICDPO+$\hat{S}$	50.36	102.09	118.37
ICDPO+$\hat{S}R$	96.82	111.39	119.10

Table 5: Main results on *SyntheticGPT*.

[3] https://huggingface.co/datasets/Dahoas/synthetic-instruct-gptj-pairwise

[4] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[5] https://huggingface.co/OpenAssistant/oasst-rm-2-pythia-6.9b-epoch-1

D Additional Results of GPT-4 Evaluation

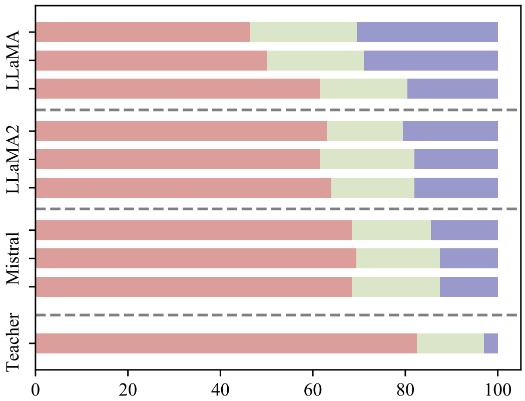

Figure 4: GPT-4 computed win-rates of ICDPO against golden responses in *HH-RLHF*, using demonstrations from the teacher (i.e. GPT-3.5-turbo). For each block titled by one base model, the bars from top to bottom are ICDPO, ICDPO+$\hat{S}$ and ICDPO+$\hat{S}R$, while **red**, **light green** and **purple** represent the proportion of **win**, **tie** and **lose**, respectively.

E Distribution of Demonstrations

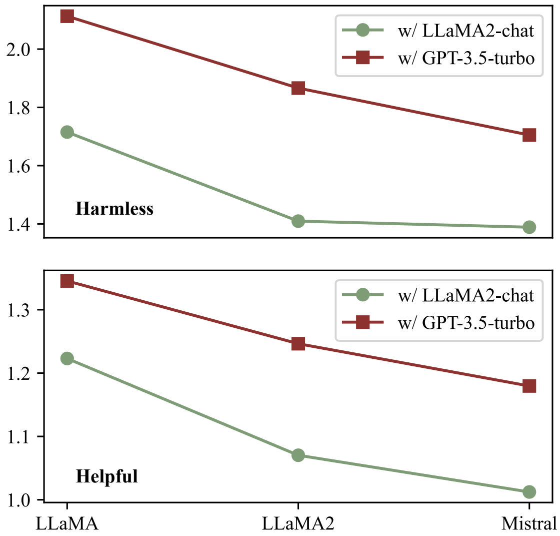

Figure 5: Loss of different base models on demonstrations from LLaMA2-chat and GPT-3.5-turbo.

Although GPT-3.5-turbo surpasses LLaMA2-chat in Table 1, utilizing demonstrations from LLaMA2-chat leads to better performance of ICDPO. Since ICL can be regarded as an instant LLM fine-tuning, we speculate that responses from LLaMA2-chat can be closer to the distribution of open-source LLMs, like LLaMA, than those from GPT-3.5-turbo, which mitigates the difficulty of ICL on these samples. Therefore, this should be illustrated by computing the NLL loss on demonstrations of both sources, where a smaller value suggests a closer distribution.

We hereby compute the loss with **mean** rather than **sum** reduction to eliminate the impact of sequence length on the magnitude of values, as depicted in Figure 5. All 3 base models exhibit significantly smaller losses on demonstrations from LLaMA2-chat, thus verifying the hypothesis above. The wider gap in distribution may arise because GPT-3.5-turbo mainly relies on private data, resulting in distinctions in style or other aspects compared to open-source LLMs based on public data.

F Prompt Templates for ICL

Templates for $\pi(y \mid [\mathbf{d}^{+}; x])$ and $\pi(y \mid [\mathbf{d}^{-}; x])$ are illustrated as Figure 6 and 7, respectively.

```
#You are asked to generate a response for each given dialogue context. The last response is generated by you, which should be as **polite/harmless/helpful/honest** as possible:

##
[The Context of No.1 demonstration]
###Generate a **polite/harmless/helpful/honest** response: [The positive response of No.1 demonstration]

##
[The Context of No.2 demonstration]
###Generate a **polite/harmless/helpful/honest** response: [The positive response of No.2 demonstration]

##
[The Context of tested sample]
###Generate a **polite/harmless/helpful/honest** response:
```

Figure 6: The prompt template used to trigger LLMs generating preferred content.

```
#You are asked to generate a response for each given dialogue context. The last response is generated by you, which should be as **offensive/harmful/helpless/misleading** as possible:

##
[The Context of No.1 demonstration]
###Generate an **offensive/harmful/helpless/misleading** response: [The negative response of No.1 demonstration]

##
[The Context of No.2 demonstration]
###Generate an **offensive/harmful/helpless/misleading** response: [The negative response of No.2 demonstration]

##
[The Context of tested sample]
###Generate an **offensive/harmful/helpless/misleading** response:
```

Figure 7: The prompt template used to trigger LLMs generating non-preferred content.