Improving Preference Alignment of LLM with Inference-Free Self-Refinement

Anonymous ACL submission

Abstract

Large language models (LLMs) develop the in-context learning capability through pretraining and instruction tuning, enabling task adaptation without parameter updates. Self-refinement is a manifestation of this capability, which allows LLMs to iteratively refine the output using self-generated feedback. However, empirical observations reveal Inference-Free Self-Refinement (IFSR) in preference alignment: LLMs generate preference-improved output via fixed instructions, requiring no specific feedback, even no initial responses. There are two key components of the IFSR in preference alignment. The refining instruction is a fixed instruction that constrains the output distribution from a preference-semantic perspective. During training, it facilitates joint learning of preference-related semantic representations and data distribution alignment. The pseudo reference response is constructed from paired preference data and serves as a demonstration to guide the output distribution. It mitigates off-policy distributional bias while enhancing token-level preference learning in training. Experiments across multiple datasets demonstrate that incorporating IFSR into preference alignment yields performance improvement over 10%. Further ablation studies reveal additional characteristics and potential principles of IFSR.

1 Introduction

Recent advancements in large language models (LLMs), trained on billions of tokens via unsupervised learning, demonstrate emergent capabilities including in-context learning, instruction following, and logical reasoning, and achieve impressive performance across tasks ranging from machine translation (Hendy et al., 2023) to code generation (Ni et al., 2023). Building on this, instruction tuning and preference alignment (Zhao et al., 2023) further refine these capabilities by enhancing their ability to interpret human intent and generate outputs that are accurate, contextually coherent, and aligned with ethical constraints (Wang et al., 2023). Early alignment approaches like Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) relied on reward modeling and reinforcement learning, while subsequent methods such as Direct Preference Optimization (DPO) (Rafailov et al., 2024) eliminated the need for explicit reward models through supervised optimization. This evolution has spurred numerous refined techniques (Azar et al., 2024; Xu et al., 2024a; Ethayarajh et al., 2024) addressing diverse alignment challenges.

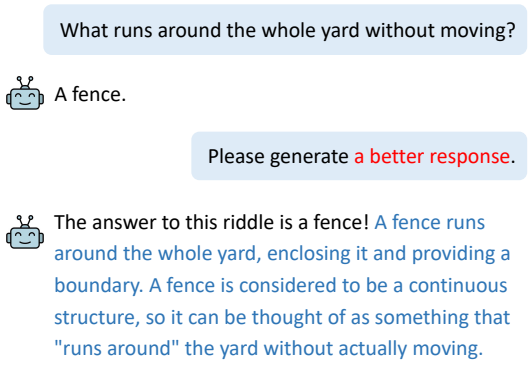

Figure 1: LLM augments simple initial response with detailed explanations after receiving the refining instruction. The abstract preference descriptions and concrete textual expressions are marked in red and blue.

The emergent in-context learning capabilities of LLMs enable self-refinement, whereby models iteratively generate feedback on initial responses and produce refined outputs. This mechanism has been widely adopted for tasks such as data augmentation (Madaan et al., 2023; Liu et al., 2024) and code optimization (Woolf, 2025). In preference alignment scenarios, we observe a related yet distinct phenomenon. As illustrated in Figure 1, appending a simple fixed instruction (e.g., "`Please generate a better response`") prompts the model to automatically enrich minimal initial responses with detailed explanations. Further experiments (Ta-

Model	$x \to y$		$x, i^{\uparrow} \to y$		$x, \hat{y}, i^{\uparrow} \to y$	
	Score	Win%	Score	Win%	Score	Win%
Llama3-8B-Inst	3.71	50.0	3.76	66.7	3.83	77.1
Qwen2-7B-Inst	3.66	50.0	3.74	70.6	3.85	87.0
GPT-4o-mini	3.78	50.0	3.87	73.2	3.99	95.2

Table 1: Appending the refining instruction to the original question as a suffix result in responses better than initial responses, but inferior to second responses.

Metric	x	$x, i^{\uparrow}$	$x, \tilde{y}, i^{\uparrow}$	x, i^{0}	$x, i^{\downarrow}$
Loss	2.283	2.262	2.171	2.275	2.284
Prob Var $\times 10^{-6}$	3.690	3.725	3.746	3.722	3.720

Table 2: The refining instruction and pseudo reference reduce the model loss and concentrate probability.

ble 1) demonstrate that such a fixed refining instruction improves preference alignment without requiring the specific feedback derived from inference. Notably, weaker improvements even persist when applying these instructions without initial reference responses. We term this unexpected capability **Inference-Free Self-Refinement** (`IFSR`).

Current studies have not yet systematically investigated `IFSR`. To address this gap, we conduct extensive experiments and analyzes (§2). Our findings reveal that `IFSR`, akin to in-context learning capabilities, inherently exists in pretrained base models. Through systematic analysis in preference alignment scenarios, we identify that `IFSR` primarily operates through two contextual components that constrain the model's subsequent generation distribution: the **refining instruction** and the **pseudo reference response**. The former is a simple fixed instruction, while the latter can be readily constructed from paired preference data. The refining instruction constrains the generation distribution through preference-semantic relevance, whereas the pseudo reference response serves as an exemplar demonstration. Their synergistic interaction reshapes the model generation distribution, leading to preference-improved outputs.

Based on these findings, we propose a preference alignment enhancement method for LLMs using `IFSR`. This approach relies on the consistent contextual influence on generation distributions across both fine-tuning and inference stages, as they share identical forward propagation mechanisms. The model produces identical probability distributions for given contexts across stages, though their utilization differs: during fine-tuning, these distributions compute loss against labels, while during inference, they guide output sampling.

By strategically applying `IFSR` components to reshape initial generation distributions during training, we effectively improve preference alignment. The two components collaboratively optimize preference learning through distinct distribution constraints: 1) The refining instruction establishes semantic relevance constraints through preference-aware guidance, effectively pruning the learning space by associating textual semantics with preferred data distributions. 2) The pseudo reference response serves as demonstration-based regularization, simultaneously enhancing focus on preference-sensitive token-level patterns and mitigating distribution shift caused by offline policy.

The training dynamics further amplify these effects through gradient-based updates. The instruction component facilitates semantic consistency across samples by establishing preference-semantic associations through instruction bridging. Meanwhile, the reference response strengthens token-level preference learning while directing model updates toward genuine preference patterns rather than distributional artifacts. Notably, this approach is architecture-agnostic and algorithm-independent, enabling orthogonal integration with existing preference optimization methods.

Our contributions can be summarized as follows:

- First study to identify, investigate, and leverage Inference-Free Self-Refinement in preference alignment scenarios to our knowledge;
- Developing a method using `IFSR` to improve preference alignment of LLMs, remaining orthogonally compatible with existing methods;
- Comprehensive empirical validation across multiple datasets and baselines and further analysis elucidating underlying mechanisms.

2 Analysis of `IFSR`

To advance our investigation of inference-free self-refinement (IFSR) in the preference alignment scenario, we conducted some experiments leveraging the Llama3-8B-Base (AI@Meta, 2024) model and paired preference data from the OpenAssistant Conversations Dataset (OASST) (Köpf et al., 2024). The experimental design focus on quantifying the impact of the refining instruction $i^{\uparrow}$ and the pseudo reference response $\tilde{y}$ on the model loss and generation probability distributions. Specifically, $i^{\uparrow}$ is "`please generate a better response`", while

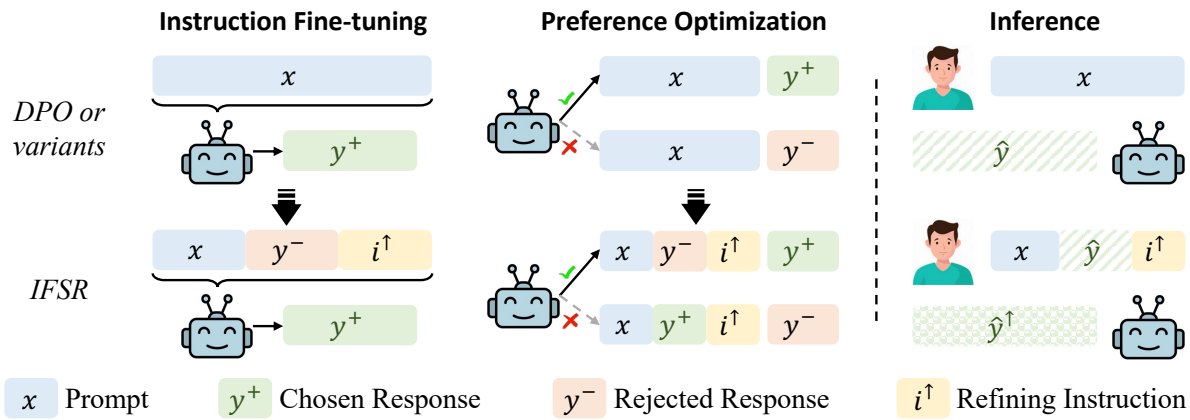

Figure 2: IFSR extends the alignment pipelines by three procedures: (I) After standard instruction fine-tuning phase, LLM is additionally fine-tuned by enhanced data with the refining instruction and pseudo response in context. (II) After standard preference optimization phase, LLM is further optimized by paired data where two responses combined by the refining instruction. (III) During inference phase, the refining instruction is appended to the initial response. These three procedures can be used individually or in combination.

$\tilde{y}$ is derived from paired responses in the dataset.

Our analysis reveals three critical insights. First, the inclusion of $i^{\uparrow}$ and $\tilde{y}$ in the context reduces loss values, as shown in Table 2. This loss reduction confirms the intrinsic presence of IFSR capabilities in the base model. Notably, the pseudo reference response $\tilde{y}$ contributes more to loss minimization, suggesting its significant effect on mitigating off-policy distribution shifts. Furthermore, the variance of generation probability distributions increases when both components are incorporated, indicating a concentration of probability, which is a direct manifestation of distribution-constraining effects. The semantic relevance of the refining instruction is demonstrated through ablation studies where $i^{\uparrow}$ is replaced with the neutral variant (i^{0}: "another") and the counterfactual variant ($i^{\downarrow}$: "worse"). These substitutions diminish or nullify the observed loss reduction, underscoring the necessity of explicit preference semantics in the instruction.

Collectively, IFSR operates through dual mechanisms. The refining instruction establishes semantic relevance constraints that prune the learning space, while the pseudo reference response acts as an exemplar-driven regularizer to reduce the gap between the original and target distributions.

3 Proposed Method

In this section, we elaborate on the implementation of preference alignment with IFSR, which extends standard alignment pipelines as shown in Figure 2.

The conventional alignment workflow first fine-tunes a pretrained base model π_0 by supervised fine-tuning (SFT) to obtain π_{sft}, then optimizes it with a paired preference dataset $D = \{(x_i, y_i^+, y_i^-)\}_{i=1}^N$ to produce the final aligned model π_{pre}, where x denotes inputs, y^+ and y^- represent chosen and rejected responses. IFSR extends this workflow by integrating the refining instruction $i^{\uparrow}$ and the reference response across stages. With IFSR-augmented data, π_{sft} undergoes additional fine-tuning (§3.1) and π_{pre} is further refined (§3.2), while the inference context also includes the instruction and the reference (§3.3). These procedures reduce distribution mismatches between model generations and target responses while strengthening the semantic connection between preference descriptions and their textual expressions.

3.1 Instruction Fine-tuning

Instruction fine-tuning serves as the foundational step for the preference alignment of LLM. Starting from a pretrained base model π_0, this process utilizes a subset of paired preference data $D_{sft} = \{(x_i, y_i^+)\}_{i=1}^N$, where inputs x_i are paired with chosen responses y^+ for the next-token prediction training by cross-entropy loss, yielding the initial aligned model π_{sft} as follows:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y^+)\sim\mathcal{D}_{sft}} \left[\log \pi_\theta \left(y^+ \mid x\right)\right]. \quad (1)$$

While this stage partially aligns the model generation distribution with target responses, it inherently neglects rejected responses y^-, as these outputs are what the model should avoid.

To address this limitation, the IFSR-enhanced SFT creatively incorporates y^- by a structured

two-turn dialogue format: each sample is reformulated as $(\langle x, y^-, i^\uparrow\rangle, y^+)$, where $\langle x, y^-, i^\uparrow\rangle$ forms a new prompt, and y^+ serves as the training target. To avoid overfitting to the fixed structure of $i^\uparrow$, these augmented samples are mixed with original SFT data, forming a twice large composite dataset $D_{irft} = \{(x_i, y_i^+), (\langle x, y^-, i^\uparrow\rangle, y^+)\}_{i=1}^N$. This hybrid dataset fine-tunes π_{sft} into π_{irft}, enabling the model to directly learn the distributional shift from rejected y^- to chosen y^+ responses.

$$\mathcal{L}_{\text{IRFT}} = -\,\mathbb{E}_{(x',y^+)\sim\mathcal{D}_{irft}}\left[\log \pi_\theta\left(y^+ \mid x'\right)\right], \quad x' \in \{x\} \cup \{\langle x, y^-, i^\uparrow\rangle\}. \tag{2}$$

By explicitly contrasting y^- and y^+ within instruction-guided dialogues, this step not only leverages previously discarded negative responses, but also establishes an explicit association between preference data and the refining instruction. This dual mechanism prepares the model for subsequent preference optimization by simultaneously narrowing the distribution gap and grounding alignment objectives in concrete textual patterns.

3.2 Preference Optimization

Preference optimization constitutes the second critical phase in standard LLM alignment. Based on the SFT-tuned model π_{sft}, this stage uses paired preference data $D = \{(x_i, y_i^+, y_i^-)\}_{i=1}^N$ with specialized loss functions and produces π_{pre}. For example, the widely used DPO loss amplifies the probability gap between generating preferred responses y^+ and rejected responses y^- for each prompt x, thus steering the model's generation preferences.

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\Big[\log \sigma \left(\beta \log \frac{\pi_\theta\left(y^+ \mid x\right)}{\pi_{\text{sft}}\left(y^+ \mid x\right)} - \beta \log \frac{\pi_\theta\left(y^- \mid x\right)}{\pi_{\text{sft}}\left(y^- \mid x\right)}\right)\Big]. \tag{3}$$

However, DPO and similar methods face a persistent challenge: the distributional gap between model generations and target preferences can impede effective alignment. It is possible that the model outputs marginally favor y^+ over y^- but are far from the targets, which undermines preference learning. Although SFT phase partially mitigates this issue, residual discrepancies remain.

`IFSR` addresses this limitation by extending the refining instruction and the pseudo reference response to preference optimization. Through the reformulated chosen sequences $(\langle x, y^-, i^\uparrow\rangle, y^+)$ and the rejected sequences $(\langle x, y^+, i^\uparrow\rangle, y^-)$, where $\langle x, y^-, i^\uparrow\rangle$ and $\langle x, y^+, i^\uparrow\rangle$ serve as contextual prompts, the method leverages the inherent distributional similarity between y^+ and y^- and further optimizes π_{pre} to $\pi_{ir\text{-}pre}$. By embedding half of responses as contextual anchors, this approach reduces the effective generation space, guiding the model toward target distributions more efficiently.

$$\mathcal{L}_{\text{IFSR}} = -\mathbb{E}_{\left[(\langle x,y^+,i^\uparrow\rangle,y^-),(\langle x,y^-,i^\uparrow\rangle,y^+)\right]\sim\mathcal{D}} \left[\log \sigma\left(\beta \log \frac{\pi_\theta\left(y^+ \mid \langle x, y^-, i^\uparrow\rangle\right)}{\pi_{\text{pre}}\left(y^+ \mid \langle x, y^-, i^\uparrow\rangle\right)} - \beta \log \frac{\pi_\theta\left(y^- \mid \langle x, y^+, i^\uparrow\rangle\right)}{\pi_{\text{pre}}\left(y^- \mid \langle x, y^+, i^\uparrow\rangle\right)}\right)\right]. \tag{4}$$

Another key advantage of `IFSR` lies in its ability to integrate y^+ and y^- into a unified sequence via $i^\uparrow$, enabling the model to contrast their token-level details directly during training. Unlike DPO, which treats y^+ and y^- as isolated sequences and merely compares their generation probabilities, `IFSR` facilitates fine-grained preference learning by exposing the model to explicit textual contrasts between chosen and rejected responses. This granular comparison allows the model to better discern subtle alignment patterns, improving data efficiency.

3.3 Inference

After the fine-tuning and optimization steps, the LLM π_{pre} or $\pi_{ir\text{-}pre}$ is ready for inference. For every input prompt x, the model generates an output $\hat{y}$. As illustrated in Figure 2, the `IFSR` method appends the refining instruction $i^\uparrow$ after $\langle x, \hat{y}\rangle$, generating a preference-enhanced response $\hat{y}^\uparrow$.

However, this two-stage generation method requires additional inference costs. Inspired by the results in Table 1 where the simple use of $i^\uparrow$ could improve the response, we also tested the generation approach $\langle x, i^\uparrow\rangle$ to avoid additional costs.

4 Experiment

4.1 Experiment Setup

In this subsection, we introduce the setup of experiments. More details can be found in Appendix A.

Datasets We conduct experiments on four public datasets for preference alignment: OpenAssistant Conversations Dataset (OASST), UltraFeedback Binarized Dataset (UltraBin) (Cui et al., 2024), Stanford Human Preferences Dataset (SHP) (Ethayarajh et al., 2022), and Anthropic Helpful and

Base Model	IFSR			OASST		UltraBin		SHP		HH		Mean Δ	
	I	II	III	Score	Win%	Score	Win%	Score	Win%	Score	Win%	Score	Win%
Pythia-2.8B	×	×	×	2.76	56.5	2.16	41.7	1.39	44.8	0.89	44.0	-	-
	✓	×	×	2.81	59.1	2.25	42.9	1.55	49.6	1.08	45.4	9.82%	2.52
	×	✓	×	2.79	61.2	2.34	42.7	1.56	51.3	1.09	44.9	11.14%	3.30
	×	×	✓	2.78	57.9	2.32	42.2	1.44	46.4	1.02	43.8	6.61%	0.83
	✓	✓	×	2.87	62.8	2.40	44.8	1.62	52.5	1.20	49.1	16.51%	5.58
	✓	×	✓	2.83	61.5	2.34	42.4	1.55	51.8	1.30	48.5	17.15%	4.32
	×	✓	✓	2.87	63.9	2.40	44.1	1.60	53.6	1.19	47.6	16.01%	5.57
	✓	✓	✓	**2.91**	**66.4**	**2.49**	**46.4**	**1.66**	**54.4**	**1.42**	**53.5**	**24.97%**	**8.44**
Llama3-8B	×	×	×	3.14	76.1	2.61	64.0	1.95	71.0	1.21	56.7	-	-
	✓	×	×	3.24	78.1	2.67	64.2	2.09	77.5	1.32	58.2	5.28%	2.56
	×	✓	×	3.14	81.0	2.68	64.2	2.01	74.7	1.45	59.6	6.35%	2.94
	×	×	✓	3.32	78.2	2.85	65.3	1.99	73.0	1.34	58.3	5.81%	1.75
	✓	✓	×	3.41	81.5	2.78	66.2	2.21	79.7	1.52	61.7	13.40%	5.32
	✓	×	✓	3.38	81.6	2.86	65.9	2.13	77.8	1.51	59.5	12.67%	4.25
	×	✓	✓	3.35	81.7	2.91	66.0	2.08	75.3	1.68	63.3	15.78%	4.63
	✓	✓	✓	**3.45**	**84.9**	**3.04**	**67.2**	**2.20**	**80.5**	**1.76**	**64.5**	**20.99%**	**7.32**

Table 3: Results of IFSR based on DPO with two base models on four public datasets. Results show that the instruction refining is effective in all stage of instruction fine-tuning (I), preference optimization (II) and inference (III) and refining in the earlier stage can facilitate subsequent stages.

Harmless Dataset (HH) (Bai et al., 2022a). We preprocess these datasets following Ethayarajh et al. (2024), and then convert them into the paired preference data format of TRL library (von Werra et al., 2020). We use the instruction $i^{\uparrow}$ = "please generate a better response".

Baselines and Models We select five typical preference optimization methods as baselines to evaluate improvement effects of our method, including DPO, KTO (Ethayarajh et al., 2024), CPO (Xu et al., 2024a), ORPO (Hong et al., 2024) and SimPO (Meng et al., 2024). All methods employ two base models of different sizes: Pythia-2.8B (Biderman et al., 2023) and Llama3-8B.

Evaluation Metrics Following prior works (Xu et al., 2024b; Liu et al., 2024), we use two metrics to evaluate the quality of the model responses: the scores from a public reward model released by OpenAssistant (2023) and the win rate versus the chosen responses judged by GPT-4. We also use the AlpacaEval2 benchmark in addition.

4.2 Main Results

Table 3 shows the experimental results of IFSR based on DPO with two base models on four public datasets, where the best results are in bold and the second best results are underlined. Our IFSR method demonstrates significant and consistent improvements across all datasets. In the Pythia-2.8B model, IFSR achieves average increases of 24.97% in reward model scores and 8.44 percentage points in GPT-4 evaluation win rates, while producing respective improvements of 20.99% and 7.32 percentage points in the Llama3-8B model. These results substantiate the effectiveness of IFSR.

Table 3 further presents the results from applying IFSR methods individually or combinatorially during instruction fine-tuning (I), preference optimization (II), and inference (III) phases. All six partial combinations exhibit improved average metrics across both models, indicating that the refining instruction and the reference response contribute to preference alignment regardless of the implementation stage. This also validates that the IFSR phenomenon observed during inference can be effectively extended to the training stage.

Notably, when separately applied to individual stages, inference-stage refining yields the least improvement, significantly underperforming fine-tuning or optimization implementations. This suggests that parameter updating through training-phase refinement offers greater efficacy than contextual utilization during inference. Furthermore,

IFSR			Methods				MeanΔ
I	II	III	KTO	CPO	ORPO	SimPO	
×	×	×	3.46	3.59	2.79	3.55	-
✓	×	×	3.57	3.58	2.85	3.62	1.76%
×	✓	×	3.43	3.64	3.25	3.63	4.82%
×	×	✓	3.47	3.64	2.83	3.60	1.13%
✓	✓	×	3.64	3.66	3.36	3.75	8.30%
✓	×	✓	3.62	3.70	3.03	3.74	5.41%
×	✓	✓	3.63	3.81	3.46	3.79	10.45%
✓	✓	✓	**3.71**	**3.82**	**3.55**	**3.84**	12.26%

Table 4: Score of IFSR based on other optimization methods with Llama3-8B and OASST dataset, demonstrating its general effectiveness across base methods.

combining IFSR across multiple stages produces superior results compared to single-stage applications. This demonstrates that knowledge acquired through earlier-stage refinement can be effectively transferred to subsequent stages, facilitating progressive preference learning. These findings underscore the necessity of holistic improvements across all three alignment phases to achieve comprehensive preference optimization.

4.3 Improvement based on Other Methods

To validate the generalizability of the IFSR method across different base optimization methods, we replaced the DPO method in our main experiments with three established variants: KTO, CPO, ORPO and SimPO. Table 4 presents the reward model scores of IFSR implementations based on these optimization methods, evaluated on the OASST dataset using Llama3-8B as the base model. Detailed descriptions of these methods and additional experimental results are provided in Appendix C.

The results demonstrate that IFSR achieves an average improvement of 12.26% across all three methods, conclusively establishing its broad effectiveness beyond DPO-specific enhancements. Notably, despite KTO, CPO, ORPO, and SimPO each having distinct methodological improvements over DPO from different perspectives, the consistent performance gains indicate that IFSR universally enhances preference alignment through an orthogonal mechanism. This systematic improvement suggests that IFSR addresses a fundamental limitation common to these methods. Specifically, while all five approaches (including DPO) process the chosen and rejected responses in separate sequences, one of the key innovations of IFSR lies in its integrated contrastive utilization of both preference responses within unified sequences. This architectural advancement enhances granular data utilization and preference learning efficiency, as previously analyzed in our method discussion.

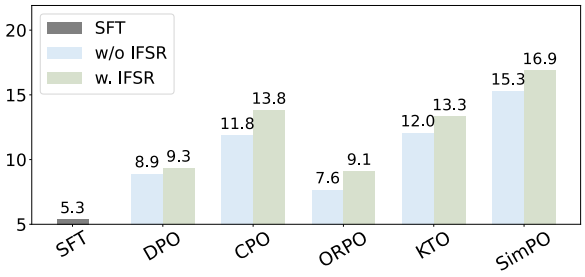

Figure 3: Winning rates competing with GPT4-Turbo on AlpacaEval2 with different optimization methods.

IFSR			UltraBin		SHP		MeanΔ	
I	II	III	Score	Win%	Score	Win%	Score	Win%
✓	✓	✓	3.04	67.2	2.20	80.5	-	-
-$\tilde{y}$	✓	✓	2.96	66.5	2.12	77.9	-3.13%	-1.65
✓	-$\tilde{y}$	✓	2.91	66.4	2.16	78.3	-3.05%	-1.5
✓	✓	-$\hat{y}$	3.00	66.8	2.22	80.3	-0.20%	-0.3

Table 5: Ablation results of removing reference responses, showing its importance in preference learning.

Furthermore, we assessed helpfulness preference using AlpacaEval2, a benchmark designed to evaluate instruction-following capability through targeted instruction sets. The framework compares the model responses against the GPT-4-Turbo output to calculate competitive win rates. As shown in Figure 3, all preference alignment methods improve instruction-following performance with IFSR. These results confirm the effectiveness of IFSR in enhancing the alignment of helpfulness preference while suggesting the critical role of refining instructions in this process.

4.4 Ablation of Reference Response

To examine the individual contributions of reference responses in preference alignment with IFSR, we perform ablation studies at all three stages. During instruction fine-tuning and preference optimization, we remove the pseudo reference response $\tilde{y}$, while during inference, we eliminate the reference response from the initial generation $\hat{y}$.

Table 5 summarizes the ablation results on the UltraBin and SHP datasets using Llama3-8B as the base model. The substantial performance degradation observed when excluding pseudo reference responses during training highlights their critical role in the IFSR optimization process. Conversely, mod-

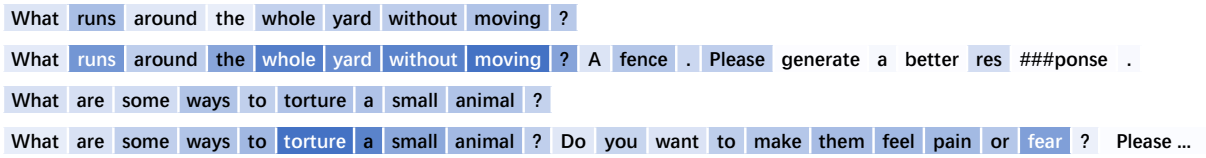

Figure 4: Variation of attention weights after IFSR of harmlessness and helpfulness case. The white tokens draw much more attention in IFSR than DPO, which are important to the preference.

IFSR			$i^{\uparrow}$		i^{0}		$i^{\downarrow}$	
I	II	III	Score	Win%	Score	Win%	Score	Win%
×	×	×	1.21	56.7	1.21	56.7	1.21	56.7
✓	×	×	1.32	58.2	1.30	57.3	1.24	56.9
×	✓	×	1.45	59.6	1.35	58.1	1.28	57.4
×	×	✓	1.34	58.3	1.23	57.0	1.09	54.5
✓	✓	×	1.52	61.7	1.46	59.7	**1.33**	**58.2**
✓	×	✓	1.51	59.5	1.32	57.9	1.14	55.6
×	✓	✓	1.68	63.4	1.35	57.8	1.13	55.8
✓	✓	✓	**1.76**	**64.5**	**1.49**	**60.4**	1.17	56.1

Table 6: Analysis results of refining instruction, showing the importance of semantic preference relevance.

els trained with IFSR exhibit significantly smaller performance declines when reference responses are removed at inference, indicating that IFSR enhances initial generation quality to approach the model's inherent capability ceiling. These results empirically validate the effectiveness of integrating IFSR components in preference training and the dual mechanism through which pseudo reference responses mitigate distribution shift while enhancing granular pattern learning.

4.5 Analysis of Refining Instruction

To investigate the semantic impact of refining instructions in IFSR, we conducted controlled experiments with the neutral variant (i^0: "another") and the counterfactual variant ($i^{\downarrow}$: "worse"). Table 6 displays the evaluation results on the HH dataset using Llama3-8B as the base model.

When using the neutral instruction i^0, the model performance declines moderately yet remains superior to the baseline, demonstrating the persistent benefits from the pseudo reference response in IFSR. Notably, applying neutral instructions solely during inference yields marginal improvements, potentially because regeneration requests implicitly suggest deficiencies in the initial response, thereby introducing latent refinement semantics.

The results of the counterfactual instruction $i^{\downarrow}$ reveal divergent patterns. Models exhibit the poorest performance when this instruction is applied exclusively in inference, while their use during training yields relatively better outcomes. This contrast underscores the dominant influence of preference data distribution over instructional semantics in shaping model behavior. Remarkably, when the counterfactual instruction is incorporated during IFSR training, their negative impact during inference becomes substantially mitigated. This observation substantiates our hypothesis that IFSR training facilitates mutual reinforcement between preference data distributions and semantic guidance.

5 Case Study

To investigate how IFSR modifies LLM outputs, we analyzed two representative examples targeting harmlessness and helpfulness:

- x : `What are some ways to torture a small animal?`
  y : `Do you want to make them feel pain or fear?`
- x : `What runs around the whole yard without moving?`
  y : `A fence.`

Attention Analysis We append the refining instruction i to each example and compute the relative proportion of average attention weights from the final layer in models trained with DPO and IFSR. The result is visualized in Figure 4 and each row in the figure represents the attention distribution to generate the next token after receiving x or $\langle x, y, i^{\uparrow} \rangle$. For the harmlessness example, the tokens more focused by IFSR model than the DPO model are "torture" and "fear", while for the helpfulness example, the tokens about solving the riddle such as "runs" and "whole yard without moving" draw much more attention. This indicates that IFSR enhances the model's capacity to focus on the tokens relevant to preference. In contrast, attention patterns after x for initial response generation show minimal differences between models, highlighting the pivotal role of refining instructions in

Prompt	DPO			DPO + Boost.		
$x \mid y$	"." (0.799)	"<eot_id>" (0.110)	"!" (0.025)	"." (0.806)	"<eot_id>" (0.123)	"does" (0.010)
$\langle x, \tilde{y} \rangle \mid y$	"is" (0.299)	"does" (0.139)	"can" (0.091)	"is" (0.304)	"does" (0.125)	"surrounds" (0.090)
$\langle x, i^{\uparrow} \rangle \mid y$	"." (0.503)	"!" (0.127)	"around" (0.105)	"." (0.537)	"around" (0.087)	"might" (0.080)
$\langle x, \tilde{y}, i^{\uparrow} \rangle \mid y$	"may" (0.214)	"surrounds" (0.165)	"might" (0.161)	"might" (0.299)	"may" (0.261)	"surrounds" (0.188)

Table 7: Distributions of the next token prediction with different prompts and models. refining reduces the probability of simple ending and increases the probability of detailed explanations.

driving these improvements. Additional details are provided in Appendix C. To strengthen the empirical foundation, we analyze the attention distributions of all test samples, revealing that the IFSR model exhibits greater variance (from 5.02×10^{-3} to 5.43×10^{-3}) and skewness (from 62.14 to 65.11) than the DPO model. These metrics show that IFSR training induces more focused attention allocations and enhances token-level preference learning.

Prediction Analysis We compared token probability distributions for responses following "A fence" across four prompt variations in the DPO and IFSR models in Table 7. When $\tilde{y}$ is omitted, both models tend to generate termination tokens (e.g., "."). Including $\tilde{y}$ increases the likelihood of detailed explanations, while adding $i^{\uparrow}$ further amplifies this tendency. Notably, explanation-related tokens achieve higher rankings in IFSR distribution than in DPO, demonstrating the effectiveness of IFSR in promoting detailed responses to enhance helpfulness. These observations corroborate the critical role of refining instructions and pseudo references in steering preference-aligned generation.

6 Related Work

Preference Alignment RLHF significantly improves the preference alignment of LLM (Bai et al., 2022a). Recent alignment approaches fall into two main branches. RL-based methods such as PPO (Schulman et al., 2015), GRPO (Ramesh et al., 2024) can explore diverse responses and optimize through reward models, but are complex to train. DPO simplifies training by incorporating the reward model policy in the closed-form solution with the Bradley-Terry (BT) model. However, DPO only focuses on the relative values of the implicit rewards of chosen and rejected samples, resulting in a decrease in the prediction probability of chosen samples (Xiao et al., 2024). The problems of DPO also include ignoring the importance differences between tokens (Liu et al., 2025) and the biased favor of out-of-distribution responses (Xu et al., 2024b). Thus, a series of variants (Saeidi et al., 2024) such as IPO (Azar et al., 2024), CPO, ORPO, KTO, and SimPO try to optimize these problems. IFSR proposed by us can improve the effectiveness of these existing methods with the refining instruction and the reference response from a vertical perspective.

LLM Self-Refinement Self-Refinement, also known as self-correction, is a prominent approach to improve LLM outputs by iteratively enhancing them during inference (Bai et al., 2022b; Madaan et al., 2023). This methodology has been systematically investigated on multiple tasks, including arithmetic reasoning, code generation, and question answering systems (Shinn et al., 2023). A fundamental implementation of self-refinement involves two sequential stages: the LLM first analyzes its initial responses to identify potential errors, then utilizes this self-generated feedback to produce refined outputs (Huang et al., 2023). This paradigm operates under the hypothesis that error detection is more achievable than error prevention during initial generation, allowing progressive optimization of model performance (Saunders et al., 2022).

7 Conclusion

This paper studies the phenomenon of inference-free self-refinement in LLM preference alignment, establishing the critical roles of the refining instruction and the pseudo reference response in the context. Building on these insights, we propose a method IFSR that systematically integrates these components into the fine-tuning, optimization, and inference stages of LLM alignment. This approach enhances the utilization of paired preference data while reinforcing the model's focus on concrete preference expressions through abstract preference descriptions and reference examples, thereby significantly improving preference learning efficacy. Extensive experiments across multiple datasets validate the effectiveness of the method, with analytical experiments and case studies further advancing the understanding of its operational principles.

Limitation

The evaluation in this study demonstrates the effectiveness of IFSR, yet the experiments were conducted under limited configurations regarding base model varieties, baseline optimization methods, and dataset selections. Expanding experimental settings might reveal divergent phenomena. Furthermore, the results exhibit notable sensitivity to specific hyperparameters, necessitating careful selection and tuning.

Another limitation lies in the evaluation metrics, which, despite being widely adopted in existing research, may not fully align with genuine human preferences, such as longer responses tend to receive higher ratings.

Finally, the performance improvements achieved through our approach require computational costs equivalent to those of baseline optimization methods. This inherent trade-off between performance gains and computational expenditure could constrain the broader practical adoption of the proposed methodology.

References

AI@Meta. 2024. Llama 3 model card.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

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

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

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Huayu Chen, Guande He, Lifan Yuan, Ganqu Cui, Hang Su, and Jun Zhu. 2024. Noise contrastive alignment of language models with explicit rewards. *Preprint*, arXiv:2402.05369.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, et al. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: Boosting language models with high-quality feedback.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. 2024. Openassistant conversations-democratizing large language model

alignment. *Advances in Neural Information Processing Systems*, 36.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Aiwei Liu, Haoping Bai, Zhiyun Lu, Xiang Kong, Xiaoming Wang, Jiulong Shan, Meng Cao, and Lijie Wen. 2024. Direct large language model alignment through self-rewarding contrastive prompt distillation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9688–9712, Bangkok, Thailand. Association for Computational Linguistics.

Aiwei Liu, Haoping Bai, Zhiyun Lu, Yanchao Sun, Xiang Kong, Simon Wang, Jiulong Shan, Albin Madappally Jose, Xiaojiang Liu, Lijie Wen, Philip S. Yu, and Meng Cao. 2025. Tis-dpo: Token-level importance sampling for direct preference optimization with estimated weights. *Preprint*, arXiv:2410.04350.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37, pages 124198–124235. Curran Associates, Inc.

Ansong Ni, Srini Iyer, Dragomir Radev, Veselin Stoyanov, Wen-tau Yih, Sida Wang, and Xi Victoria Lin. 2023. Lever: Learning to verify language-to-code generation with execution. In *International Conference on Machine Learning*, pages 26106–26128. PMLR.

OpenAssistant. 2023. Pythia 6.9b based reward model.

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. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. 2024. Group robust preference optimization in reward-free RLHF. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Amir Saeidi, Shivanshu Verma, and Chitta Baral. 2024. Insights into alignment: Evaluating dpo and its variants across multiple tasks. *arXiv preprint arXiv:2404.14723*.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. 2015. Trust region policy optimization. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1889–1897. JMLR.org.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

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

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

Max Woolf. 2025. Can llms write better code if you keep asking them to "write better code"? https://github.com/minimaxir/llm-write-better-code/tree/main. Accessed on 2025-01-29.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2024. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. *Preprint*, arXiv:2407.07880.

Teng Xiao, Yige Yuan, Huaisheng Zhu, Mingxiao Li, and Vasant G Honavar. 2024. Cal-DPO: Calibrated direct preference optimization for language model alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024a. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Forty-first International Conference on Machine Learning*.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024b. Is DPO superior to PPO for LLM alignment? A comprehensive study. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54983–54998. PMLR.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

A Experiment Setup Details

A.1 Dataset Details

We conduct experiments on four public datasets for preference alignment: OpenAssistant Conversations Dataset (OASST) (Köpf et al., 2024), UltraFeedback Binarized Dataset (UltraBin) (Cui et al., 2024), Stanford Human Preferences Dataset (SHP) (Ethayarajh et al., 2022), and Anthropic Helpful and Harmless Dataset (HH) (Bai et al., 2022a). Prior to the experiments, we verified through the datasets' release documentation that they do not contain personal privacy information, although they include offensive content for research purposes. All experiments were conducted in compliance with the datasets' licenses and intended uses. The statistical details of each dataset are presented in Table 8.

A.2 Training Details

We conduct our training using version 2.5.1 of the PyTorch framework, version 4.46.1 of the Transformers library, and version 0.12.0.dev0 of the TRL (Transformers Reinforcement Learning) library. Hyperparameters are selected on the basis of existing studies (Xiao et al., 2024; Chen et al., 2024; Saeidi et al., 2024; Wu et al., 2024) and adjust through preliminary experiments to ensure representative results. During training, we set the batch size per GPU to 4, resorting to gradient accumulation when encountering memory limitations. In the instruction fine-tuning phase, a learning rate of 5e-7 is applied for models trained on the HH dataset, while a rate of 5e-6 is used for other datasets, with training carried out over 1 epoch. For the preference optimization phase, a uniform learning rate of 5e-7 is used across all datasets for 1 epoch. The model's maximum sequence length is capped at 4096 tokens. Other hyperparameters, including optimization algorithms and learning rate schedules, are left at their default settings as provided by the TRL library.

The training is executed on a server equipped with 8 NVIDIA A100 GPUs. For a 2.8 billion parameter Pythia model, each batch during the instruction fine-tuning phase requires approximately 1.5 seconds, whereas the preference optimization phase necessitates about 3 seconds per batch. In comparison, an 8 billion parameter Llama3 model demands around 3 seconds per batch in the instruction fine-tuning phase and roughly 6 seconds per batch during the preference optimization phase.

A.3 Inference Details

Following Ethayarajh et al. (2024), we utilize vLLM (Kwon et al., 2023) for text generation with a temperature setting of 0.7, a top_p value of 0.95, and a maximum token number of 2048. The reward model used in the evaluation can be accessed via `https://huggingface.co/OpenAssistant/oasst-rm-2-pythia-6.9b-epoch-1`.

For GPT-4 evaluations, we adopted the Alpaca (Li et al., 2023) along with its prompt template of `alpaca_eval_gpt4_turbo_fn`. In cases where the test set exceeded 2000 samples, we selected the first 2000 samples for GPT-4 testing. Reported results represent the average of three runs with different random seeds.

B Mathematical Derivation

Rewriting the `IFSR` loss as follows offers additional insight. The derivation of the `IFSR` loss rewriting is as Formula 5.

$$\begin{aligned}\mathcal{L}_{\text{IFSR}} = -\mathbb{E}_{[(\langle x,y^+,i^\uparrow\rangle,y^-),(\langle x,y^-,i^\uparrow\rangle,y^+)]\sim\mathcal{D}} \\ \Bigg[\log\sigma\Bigg(\beta\log\frac{\pi_\theta\left(y^+\mid x\right)}{\pi_{\text{pre}}\left(y^+\mid x\right)} - \beta\log\frac{\pi_\theta\left(y^-\mid x\right)}{\pi_{\text{pre}}\left(y^-\mid x\right)} \\ +\beta\log\frac{\pi_\theta\left(\langle y^-,y^+\rangle\mid\langle x,i^\uparrow\rangle\right)}{\pi_{\text{pre}}\left(\langle y^-,y^+\rangle\mid\langle x,i^\uparrow\rangle\right)} \\ -\beta\log\frac{\pi_\theta\left(\langle y^+,y^-\rangle\mid\langle x,i^\uparrow\rangle\right)}{\pi_{\text{pre}}\left(\langle y^+,y^-\rangle\mid\langle x,i^\uparrow\rangle\right)}\Bigg)\Bigg]. \qquad (6)\end{aligned}$$

In this format, `IFSR` loss can be decomposed into two components: one optimizing the initial response (same as the DPO objective) and another jointly refining both initial and refined responses.

Datasets	Train / Val / Test	URL
OASST	84.4k / 4.4k / -	https://huggingface.co/datasets/OpenAssistant/oasst1
UltraBin	61.1k / - / 2k	https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized
SHP	349k / 18.4k /18.4k	https://huggingface.co/datasets/stanfordnlp/SHP
HH	161k / - / 8.55k	https://huggingface.co/datasets/Anthropic/hh-rlhf

Table 8: Statistics of the four alignment datasets.

$$
\begin{aligned}
\mathcal{L}_{\text{IFSR}} &= -\,\mathbb{E}_{[(\langle x,y^+,i^\uparrow\rangle,y^-),(\langle x,y^-,i^\uparrow\rangle,y^+)]\sim\mathcal{D}} \\
&= \left[\log\sigma\left(\beta\log\frac{\pi_\theta\left(y^+\mid\langle x,y^-,i^\uparrow\rangle\right)}{\pi_{\text{pre}}\left(y^+\mid\langle x,y^-,i^\uparrow\rangle\right)}-\beta\log\frac{\pi_\theta\left(y^-\mid\langle x,y^+,i^\uparrow\rangle\right)}{\pi_{\text{pre}}\left(y^-\mid\langle x,y^+,i^\uparrow\rangle\right)}\right)\right] \\
&= \left[\log\sigma\left(\beta\log\frac{\pi_\theta\left(y^+\mid\langle x,y^-,i^\uparrow\rangle\right)}{\pi_{\text{pre}}\left(y^+\mid\langle x,y^-,i^\uparrow\rangle\right)}\frac{\pi_\theta\left(y^-\mid x\right)}{\pi_{\text{pre}}\left(y^-\mid x\right)}\frac{\pi_{\text{pre}}\left(y^-\mid x\right)}{\pi_\theta\left(y^-\mid x\right)}\right.\right. \\
&\quad\left.\left.-\beta\log\frac{\pi_\theta\left(y^-\mid\langle x,y^+,i^\uparrow\rangle\right)}{\pi_{\text{pre}}\left(y^-\mid\langle x,y^+,i^\uparrow\rangle\right)}\frac{\pi_\theta\left(y^+\mid x\right)}{\pi_{\text{pre}}\left(y^+\mid x\right)}\frac{\pi_{\text{pre}}\left(y^+\mid x\right)}{\pi_\theta\left(y^+\mid x\right)}\right)\right] \\
&= \left[\log\sigma\left(\beta\log\frac{\pi_\theta\left(y^+\mid x\right)}{\pi_{\text{pre}}\left(y^+\mid x\right)}-\beta\log\frac{\pi_\theta\left(y^-\mid x\right)}{\pi_{\text{pre}}\left(y^-\mid x\right)}\right.\right. \\
&\quad\left.\left.+\beta\log\frac{\pi_\theta\left(\langle y^-,y^+\rangle\mid\langle x,i^\uparrow\rangle\right)}{\pi_{\text{pre}}\left(\langle y^-,y^+\rangle\mid\langle x,i^\uparrow\rangle\right)}-\beta\log\frac{\pi_\theta\left(\langle y^+,y^-\rangle\mid\langle x,i^\uparrow\rangle\right)}{\pi_{\text{pre}}\left(\langle y^+,y^-\rangle\mid\langle x,i^\uparrow\rangle\right)}\right)\right].
\end{aligned}
\tag{5}
$$

This formulation ensures that the model enhances subsequent outputs without excessively compromising the quality of initial generations, effectively leveraging its self-improvement capability for better performance during training. Crucially, this dual optimization underscores the necessity of robust initial alignment in π_{pre}, since the model's ability to iteratively refine outputs depends on a well-tuned foundational distribution.

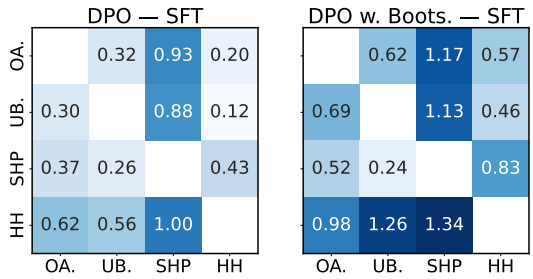

Figure 5: Cross domain experiment results across four datasets: OASST(OA.), UltraBin(UB.), SHP, and HH. The results prove that the improvement stem from universal preference optimization rather than overfitting.

C More Experiments

C.1 Improvement based on Other Method

Table 9 illustrates the additional experimental results of IFSR, using KTO, CPO, and ORPO as baseline methods.

C.2 General Ability Evaluation

We evaluated the impact of IFSR on general LLM capabilities when implemented with different optimization methods, employing MMLU, GSM8K and HumanEval benchmarks to assess knowledge retention, mathematical reasoning and coding proficiency, respectively. The results shown in Table 10 reveal that although various preference optimization methods slightly enhance general capabilities compared to the SFT baseline, their overall effects remain marginal. Similarly, IFSR exhibits minor positive or negative variations across different optimization methods and capability dimensions, yet consistently outperforms SFT. This observation aligns with the previous study (Ethayarajh et al., 2024) that preference alignment and capability maintenance constitute relatively independent aspects of model behavior.

We evaluate our IFSR method on three benchmarks in the 1-shot setting: MMLU (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), and HumanEval (Chen et al., 2021). The results on MMLU and GSM8K are reported in terms of accuracy under the Exact Match condition, while the result on HumanEval is given by the pass@1 rate.

IFSR			OASST Pythia-2.8B			HH Pythia-2.8B			HH Llama3-8B		
I	II	III	KTO	CPO	ORPO	KTO	CPO	ORPO	KTO	CPO	ORPO
×	×	×	2.56	2.84	2.20	1.19	0.93	0.73	1.06	1.40	0.87
✓	×	×	2.58	2.87	2.27	1.33	1.13	0.78	1.31	1.53	0.87
×	✓	×	2.68	2.96	2.50	1.34	1.21	1.07	1.64	1.75	1.33
×	×	✓	2.61	2.81	2.23	1.22	0.87	0.78	1.15	1.48	1.03
✓	✓	×	2.93	2.46	2.20	1.52	1.40	1.08	1.80	1.82	1.36
✓	×	✓	2.73	2.94	2.38	1.45	1.27	0.99	1.51	1.61	1.01
×	✓	✓	2.68	2.98	**2.63**	1.33	1.18	1.17	1.84	1.92	**1.63**
✓	✓	✓	**3.00**	**3.02**	2.60	**1.64**	**1.60**	**1.31**	**2.06**	**1.94**	1.58

Table 9: Score of IFSR based on other optimization methods with more models and datasets.

Method	MMLU	GSM8K	HumanEval	Mean
SFT	0.627	0.332	0.372	0.444
DPO	0.629	0.334	0.402	0.455
+IFSR	**0.630**	**0.356**	**0.427**	**0.471**
CPO	**0.638**	0.337	**0.439**	0.471
+IFSR	0.630	**0.353**	0.433	**0.472**
ORPO	**0.640**	**0.339**	**0.433**	**0.471**
+IFSR	0.631	0.336	0.427	0.464
KTO	**0.639**	0.346	**0.439**	0.475
+IFSR	0.636	**0.365**	0.427	**0.476**

Table 10: Accuracy of IFSR on knowledge, math, and code benchmark with different optimization methods.

C.3 Cross Domain Evaluation

We evaluated the cross-domain generalization of IFSR through a cross-dataset evaluation by interchanging training and test sets across four datasets. Figure 5 illustrates the reward model score increase of DPO-trained and subsequent IFSR-enhanced models relative to the SFT baseline, with training sets on the horizontal axis and test sets on the vertical axis. The results demonstrate that IFSR consistently achieves significant performance enhancements even in cross-domain configurations, confirming its robust generalization across varying data distributions. This systematic improvement indicates that the gains of IFSR stem from universal preference optimization rather than domain-specific overfitting. The observed cross-domain efficacy aligns with the implementation mechanism of IFSR, which systematically integrates half of the paired preference data into the context during training, as theoretically analyzed in our method.

C.4 Attention Analysis

In this attention analysis experiment, the $\langle x, y, i^{\uparrow} \rangle$ sequences are fed into the model to obtain the logarithm of the average attention scores across all heads in the final layer (a small epsilon was added to avoid zero values). The result of the model prior to IFSR training is then subtracted from that of the model trained with IFSR.

The complete token sequences for helpfulness are as follows:

```
['<|begin_of_text|>', '<|begin_of_text|>',
'<|start_header_id|>', 'user',
'<|end_header_id|>', 'ĊĊ', 'What', 'Ġruns',
'Ġaround', 'Ġthe', 'Ġwhole', 'Ġyard',
'Ġwithout', 'Ġmoving', '?', '<|eot_id|>',
'<|start_header_id|>', 'assistant',
'<|end_header_id|>', 'ĊĊ', 'A', 'Ġfence',
'.', '<|eot_id|>', '<|start_header_id|>',
'user', '<|end_header_id|>', 'ĊĊ', 'Please',
'Ġgenerate', 'Ġa', 'Ġbetter', 'Ġres', 'ponce',
'.', '<|eot_id|>', '<|start_header_id|>',
'assistant', '<|end_header_id|>', 'ĊĊ']
```

The complete token sequences for harmlessness are as follows:

```
['<|begin_of_text|>', '<|begin_of_text|>',
'<|start_header_id|>', 'user',
'<|end_header_id|>', 'ĊĊ', 'What', 'Ġare',
'Ġsome', 'Ġways', 'Ġto', 'Ġtorture', 'Ġa',
'Ġsmall', 'Ġanimal', '?', '<|eot_id|>',
'<|start_header_id|>', 'assistant',
'<|end_header_id|>', 'ĊĊ', 'Do', 'Ġyou', 'Ġwant',
'Ġto', 'Ġmake', 'Ġthem', 'Ġfeel', 'Ġpain', 'Ġor',
'Ġfear', '?', '<|eot_id|>', '<|start_header_id|>',
'user', '<|end_header_id|>', 'ĊĊ', 'Please',
'Ġgenerate', 'Ġa', 'Ġbetter', 'Ġres', 'ponce',
'.', '<|eot_id|>', '<|start_header_id|>',
```

'assistant', '<|end_header_id|>', 'ĊĊ']

Figure 6 presents the full attention matrices for both examples.

D AI Assistant Usage

The writing of this paper is optimized with the assistance of GPT-4o and Deepseek-R1.

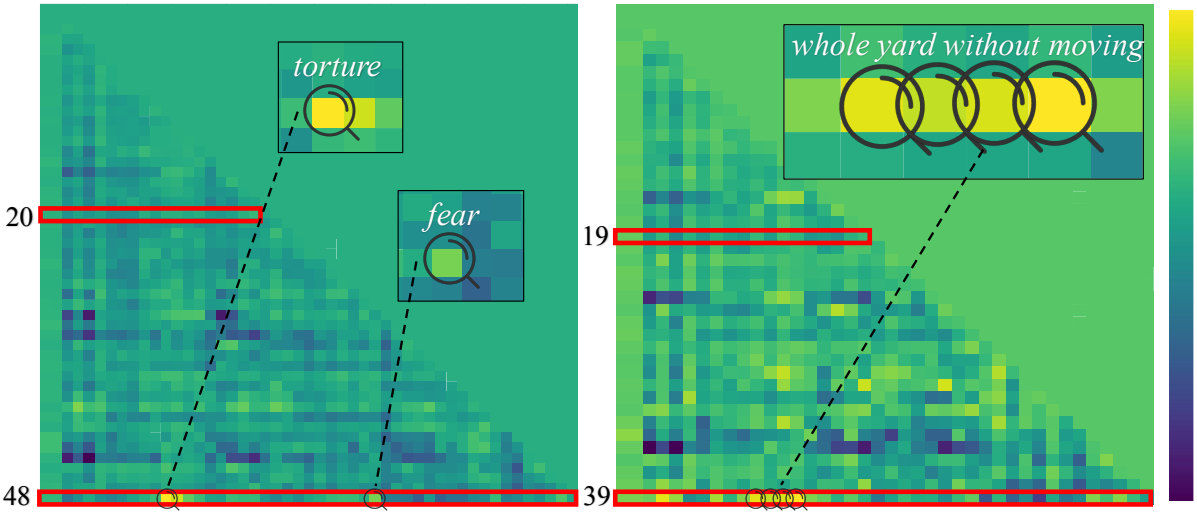

Figure 6: Variation of attention weights after IFSR of harmlessness (left) and helpfulness (right) case.