ON-POLICY REPRESENTATION ALIGNMENT IN PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct preference optimization (DPO) has become a simple and effective paradigm for aligning large language models (LLMs) with human preferences. However, DPO-style methods are fundamentally constrained by a mismatch between static offline preference data and the evolving behavior of the current policy. Recent works attempt to mitigate this mismatch by constructing preference supervision in the reward space or through data-side strategies. However, these constraints provide only a compressed view of the internal generation process and do not explicitly ensure that the current policy encodes sufficient preference information during generation. In this paper, we revisit this limitation from a representation perspective and propose On-Policy Representation Alignment (OPRA), which extends preference optimization from reward-level response ranking to representation-level preference alignment. Specifically, for an offline preference pair, we sample a short response prefix from the current policy as an anchor and encourage its response-level representation to be closer to that of the chosen response than to that of the rejected response. By jointly optimizing output-level preference consistency and on-policy representation alignment, OPRA provides a direct representation-level bridge between offline preference supervision and the policy's own autoregressive generation. Experiments on ten benchmarks demonstrate consistent improvements across preference-data settings and LLM backbones. Further analyses show that OPRA learns more preference-consistent representations and generalizes well across diverse settings.

1 INTRODUCTION

Aligning large language models (LLMs) with human preferences is essential for improving their helpfulness, safety, and instruction-following capabilities. Reinforcement Learning from Human Feedback (RLHF) has become a standard paradigm for preference alignment by learning a reward model from human comparisons and then optimizing the policy against the learned reward (Zheng et al., 2025; Gao et al., 2025; Xu & Ding, 2026; Qi et al., 2026). Despite its effectiveness, RLHF requires multiple training stages and reinforcement learning, making the overall optimization computationally expensive and sensitive to reward model and optimization instability. Direct Preference Optimization (DPO) (Rafailov et al., 2023) offers a simpler alternative by directly optimizing offline preference data without explicitly training a reward model, and has inspired broad efforts in preference optimization (Ethayarajh et al., 2024; Wu et al., 2025; Hong et al., 2024; Liu et al., 2025; Zhu et al., 2025a).

However, most DPO-style methods optimize on pre-collected offline preference pairs, whereas inference autoregressively samples responses from the current policy. Without an explicit reward model, DPO is limited in sampling preference pairs from the optimal policy. This creates a distributional gap between the policy used for data collection and the target policy (Zhou et al., 2024; Meng et al., 2024), which we refer to as the offline-to-on-policy mismatch issue. Consequently, correctly ranking offline preference pairs does not necessarily ensure preference-consistent behavior in responses generated by the current policy. Empirical study has also shown that better pairwise preference classification does not always translate into better generation quality (Tang et al., 2024).

To mitigate this offline-to-on-policy mismatch, recent methods incorporate current-policy generations into preference optimization. As shown in Figure 1 (b-c), they either reformulate reward-level

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

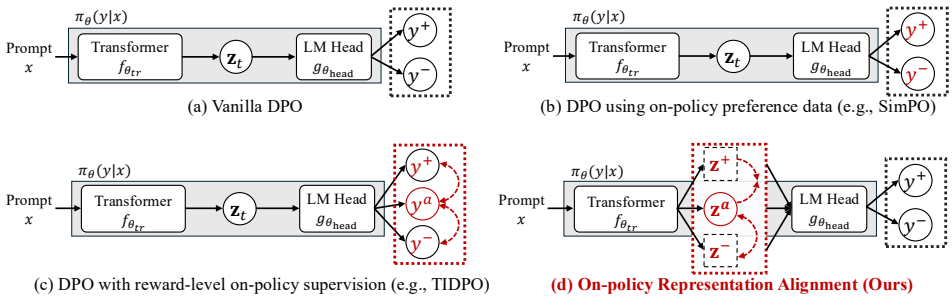

Figure 1: Comparison of OPRA with vanilla DPO and its variants for mitigating mismatch issue.

objectives (Yang et al., 2026a) or construct preference data from responses sampled by the current policy (Meng et al., 2024). However, these methods still primarily impose preference supervision through scalar reward signals derived from model outputs, which do not uniquely determine the underlying representation relations. This limitation arises from two related aspects. First, scalar rewards provide only a compressed view of the internal generation process. The LM head exposes the hidden state only through next-token probabilities, and sequence-level aggregation further collapses them into scalar preference scores. Consequently, many hidden states can induce similar sequence scores or the same pairwise reward margin, leaving their underlying representation relations underdetermined. Second, even when current-policy generations are incorporated, the offline preference relation is not directly imposed on representations induced by the current policy. Existing objectives mainly constrain policy-generated responses through scalar rewards, leaving their internal representations without explicit chosen-rejected preference supervision.

Motivated by the above challenges, this paper investigates to mitigate the offline-to-on-policy mismatch in the representation space for preference optimization, as shown in Figure 1 (d). For each offline preference pair, we sample a short response prefix from the current policy as an on-policy anchor, and then pull the anchor representation closer to the chosen response while pushing it away from the rejected response. The anchor ensures that the optimized representation is induced by the current policy, while the chosen-rejected pair defines the preference direction along which it is updated. This allows offline preference supervision to directly shape the representations arising during the policy's own autoregressive generation. Unlike existing methods that constrain current-policy responses through scalar rewards, we directly exploit the richer preference-discriminative structure of hidden representations, thereby narrowing the offline-to-on-policy mismatch.

We propose **On-Policy Representation Alignment (OPRA)** to explicitly model preference relations in representation space. For each offline preference pair, OPRA autoregressively samples a short response prefix from the current policy as an on-policy anchor and extracts response-level representations of the anchor, chosen response, and rejected response. We then introduce a preference representation alignment objective that encourages the anchor representation to be closer to the chosen target than to the rejected target. A preference-consistency gate further excludes relational targets whose current implicit rewards disagree with the annotated preference. Finally, OPRA combines this representation-level objective with DPO and chosen-response negative log-likelihood, jointly optimizing preference ranking, representation alignment, and chosen-response likelihood. Importantly, OPRA requires only a short on-policy anchor and does not require constructing and repeatedly annotating an on-policy preference dataset.

We conduct experiments on ten benchmarks covering knowledge, reasoning, generation, and instruction following. Experimental results show that OPRA consistently achieves better average performance over comparison baselines across multiple LLM backbones. Further evaluation across anchor lengths, representation layers, and alignment targets demonstrate that directly aligning hidden representations produces more preference-consistent on-policy states and generation behavior.

Our main contributions are summarized as follows: (1) We study the mismatch between offline preference optimization and online autoregressive generation from a representation-level perspective. We show that scalar rewards provide compressed preference signals and do not fully model the underlying representation relations, motivating explicit preference supervision in the representation space. (2) We propose an OPRA framework to directly transfer offline preference relations to rep-

resentations induced by the current policy. OPRA introduces a current-policy prefix as anchor, and aligns its representation with the corresponding chosen and rejected responses. (3) Experiments on ten benchmarks spanning understanding, reasoning, generation and instruction following, demonstrate the effectiveness of OPRA across multiple LLM backbones and preference datasets. Further analyses verify its generalizability across annotators, data quality, LLM scales and families.

2 PRELIMINARY

Preference Optimization We are given an offline preference dataset $\mathcal{D} = \{(x_i, y_i^+, y_i^-)\}_{i=1}^N$, where y_i^+ and y_i^- denote the chosen and rejected responses to prompt x_i, and we write ρ for the marginal distribution of prompts in $\mathcal{D}$. Under the Bradley-Terry model (Bradley & Terry, 1952), preferences are induced by a latent reward r^* through $p^*(y^+ \succ y^-|x) = \sigma\big(r^*(x, y^+) - r^*(x, y^-)\big)$, where $\sigma(\cdot)$ is the logistic function. Preference optimization seeks a policy that maximizes the expected reward while remaining close to a reference policy π_{ref}:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\rho}\Big[\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\big[r^*(x,y)\big] - \beta\, D_{\text{KL}}\big(\pi_\theta(\cdot|x) \,\big\|\, \pi_{\text{ref}}(\cdot|x)\big)\Big],$$

where $\beta > 0$ controls the strength of the KL regularization. Inverting this relation gives $r^*(x, y) = \beta \log \frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x)$, which motivates the implicit reward $r_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$.

Direct Preference Optimization (DPO) Under the Bradley-Terry preference model, DPO parameterizes an implicit reward as $r_\theta(x, y) = \beta \log[\pi_\theta(y|x)/\pi_{\text{ref}}(y|x)]$, up to a prompt-dependent additive constant (Rafailov et al., 2023). Defining $M_\theta(x, y^+, y^-) = r_\theta(x, y^+) - r_\theta(x, y^-)$ gives

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}} \log \sigma(M_\theta(x, y^+, y^-)) = -\mathbb{E}_{\mathcal{D}} \log \sigma\left(\beta \log \frac{\pi_\theta(y^+|x)}{\pi_{\text{ref}}(y^+|x)} - \beta \log \frac{\pi_\theta(y^-|x)}{\pi_{\text{ref}}(y^-|x)}\right).$$

This objective fits preference relations on the observed response pairs.

Mismatch between Offline Preference Supervision and Online Generation DPO learns preference relations $y^+ \succ y^-$ from offline response pairs $\mathcal{D} = \{(x, y^+, y^-)\}$, whereas inference generates responses from the evolving policy autoregressively. We refer to this discrepancy as the **offline-to-on-policy mismatch**. Let $p_{\mathcal{D}}(y|x)$ denote the response distribution induced by the offline preference data and $\pi_\theta(y|x)$ the current-policy distribution. The mismatch can be characterized by

$$p_{\mathcal{D}}(y|x) \neq \pi_\theta(y|x).$$

Minimizing a preference objective on $y^+ \succ y^-$ only enforces the annotated preference relation on the observed offline responses and does not directly constrain responses generated from the current policy. This motivates transferring the offline preference supervision in representation space to current-policy generations (Section 3).

3 METHODOLOGY

We first analyze why preference supervision at reward level is insufficient to provide offline preference supervision to the policy's generation, and motivate the representation-level alignment (Section 3.1). We then present the on-policy representation alignment framework and its objective (Section 3.2), and the overall objective and training algorithm (Section 3.3).

3.1 FROM REWARD-LEVEL PREFERENCE TO ON-POLICY REPRESENTATION ALIGNMENT

The policy π_θ consists of a Transformer $f_{\theta_{\text{tr}}}$ and an output head $g_{\theta_{\text{head}}}$. Let $\mathbf{h}_{\theta,t}^{(\ell)}(x, y) \in \mathbb{R}^d$ denote the hidden state of layer $\ell \in \{1, \ldots, L\}$ at the position of response token y_t. The next-token distribution is read out from the last-layer state as $\log \pi_\theta(y_t|x, y_{<t}) = \psi_{y_t}(\mathbf{h}_{\theta,t-1})$, where $\psi(\mathbf{h}) = \log \text{softmax}\big(g_{\theta_{\text{head}}}(\mathbf{h})\big)$. Hence the reward margin is a scalar function of the hidden trajectories $\mathbf{H}_\theta^\pm$ of $y^\pm$, i.e.,

$$M_\theta = \Phi\big(\mathbf{H}_\theta^+, \mathbf{H}_\theta^-\big) = \beta \sum_{t=1}^{|y^+|} \psi_{y_t^+}\big(\mathbf{h}_{\theta,t-1}(x, y^+)\big) - \beta \sum_{t=1}^{|y^-|} \psi_{y_t^-}\big(\mathbf{h}_{\theta,t-1}(x, y^-)\big) - c, \qquad (1)$$

where c refers to the reference log-probabilities.

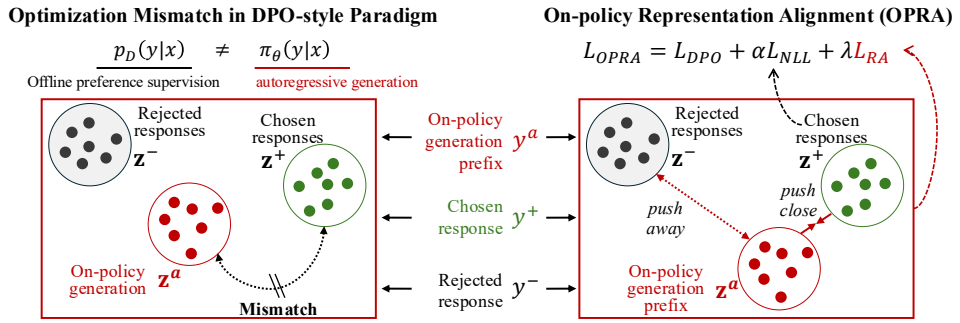

Figure 2: Overview of the proposed OPRA. Left: Offline preference supervision and current-policy autoregressive generation induce different context distributions, leaving a gap between offline preference learning and on-policy generation. Right: OPRA samples a short on-policy prefix as an anchor and encourages its response-level representation to be closer to the chosen representation than to the rejected representation. The representation alignment objective is jointly optimized with DPO and chosen-response negative log-likelihood.

Proposition 1 (Reward-level preference statistics underdetermine representation relations)**.** *Let* $\mathbf{M}_\theta \in \mathbb{R}^{N'}$ *collect scalar preference statistics such as implicit rewards or pairwise reward margins, with Jacobian* $J_M = \partial \mathbf{M}_\theta / \partial \theta$*. Let* $g_\theta \in \mathbb{R}$ *denote a differentiable relation among response representations. If* $\nabla_\theta g_\theta \notin \mathrm{row}(J_M)$*, then there exists a direction* $\mathbf{v}$ *such that*

$$\mathbf{M}_{\theta+\epsilon\mathbf{v}} = \mathbf{M}_\theta + O(\epsilon^2), \qquad g_{\theta+\epsilon\mathbf{v}} = g_\theta + \epsilon c_\mathbf{v} + O(\epsilon^2), \tag{2}$$

with $c_\mathbf{v} \neq 0$*.*

Proposition 1 shows that reward-level DPO-style objectives can be first-order insensitive to parameter directions that change representation relations while leaving the optimized scalar preference statistics unchanged. Thus, correctly optimizing implicit rewards does not uniquely determine the underlying representation structure. Constraining preference relations in representation space provides direct supervision beyond reward-level preference optimization. See Appendix B.1 for proof.

3.2 On-Policy Representation Alignment

3.2.1 On-policy Anchor and Representations

Let π_θ denote the policy model. At iteration k, for each preference pair $\xi = (x, y^+, y^-) \sim \mathcal{D}$, we sample a response prefix of at most K tokens from the current policy without gradient,

$$y^a \sim \widetilde{\pi}_{\theta_k}^{(\leq K)}(\cdot|x), \tag{3}$$

where $\widetilde{\pi}_\theta$ is the decoding distribution induced by the current policy. We denote the joint law by $Q_k(\xi, y^a) = \mathcal{D}(\xi)\, \widetilde{\pi}_{\theta_k}^{(\leq K)}(y^a|x)$. Each response is encoded by a normalized pooled representation:

$$\mathbf{z}_\theta(x,y) = \frac{\mathbf{u}_\theta(x,y)}{\|\mathbf{u}_\theta(x,y)\|_2}, \qquad \mathbf{u}_\theta(x,y) = \frac{1}{|\mathcal{I}(y)|} \sum_{t \in \mathcal{I}(y)} \mathbf{h}_{\theta,t}^{(\ell)}(x,y), \tag{4}$$

where $\ell = L$ is the last layer L of hidden states by default. The anchor and preference targets are

$$\mathbf{z}_\theta^a = \mathbf{z}_\theta(x, y^a), \qquad \bar{\mathbf{z}}_k^+ = \mathrm{sg}\big[\mathbf{z}_{\theta_k}(x, y^+)\big], \qquad \bar{\mathbf{z}}_k^- = \mathrm{sg}\big[\mathbf{z}_{\theta_k}(x, y^-)\big], \tag{5}$$

where the stop-gradient $\mathrm{sg}[\cdot]$ makes the offline pair a fixed relational target. In this way, the optimization moves the representation induced by current-policy generation while keeping the chosen and rejected targets fixed.

3.2.2 Vanilla Representation Alignment

Given a similarity function $s(\cdot,\cdot)$ in the representation space, we define the representation preference margin as

$$\Delta_{\theta,k}(\xi, y^a) = s(\mathbf{z}_\theta^a, \bar{\mathbf{z}}_k^+) - s(\mathbf{z}_\theta^a, \bar{\mathbf{z}}_k^-). \tag{6}$$

A positive margin indicates that the on-policy anchor is more similar to the chosen target than to the rejected target. We then define the vanilla representation alignment objective as

$$\mathcal{L}^{\phi}_{\mathrm{RA}^{\mathrm{van}}}(\theta;\theta_k) = \mathbb{E}_{(\xi,y^a)\sim Q_k}\left[\phi(\Delta_{\theta,k}(\xi,y^a))\right], \tag{7}$$

where $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a non-increasing preference-margin loss. Minimizing Eq. 7 encourages the representation of the policy-generated anchor to have a larger score with the chosen target than with the rejected target. Here, we consider two implementations for the representation preference margin loss function ϕ in Eq 7.

1) NCE-based RA Loss (default in experiments). We employ NCE loss (Gutmann & Hyvärinen, 2010; van den Oord et al., 2018) to implement the representation alignment. We apply the similarity function with cosine similarity. Since the response representations are ℓ_2-normalized, $s_{\mathrm{NCE}}(\mathbf{z},\mathbf{z}') = \mathbf{z}^\top\mathbf{z}'$. Treating the matched rejected response as the negative, we define

$$\ell^{\mathrm{NCE}}_{\mathrm{RA}}(\xi,y^a) = -\log\frac{\exp\big(s(\mathbf{z}^a_\theta,\bar{\mathbf{z}}^+_k)/\tau\big)}{\exp\big(s(\mathbf{z}^a_\theta,\bar{\mathbf{z}}^+_k)/\tau\big)+\exp\big(s(\mathbf{z}^a_\theta,\bar{\mathbf{z}}^-_k)/\tau\big)} = \log(1+\exp(-\Delta_{\theta,k}(\xi,y^a)/\tau))\,, \tag{8}$$

where $\tau > 0$ is the temperature. Thus, $\phi_{\mathrm{NCE}}(\Delta) = \log(1+e^{-\Delta/\tau})$.

2) Triplet-based RA Loss. We consider a margin-based triplet ranking loss that directly enforces a minimum separation between the similarities to the chosen and rejected targets. We employ a triplet loss (Schroff et al., 2015) and define $s_{\mathrm{tri}}(\mathbf{z},\mathbf{z}') = -\|\mathbf{z}-\mathbf{z}'\|_2$. Then, we have:

$$\ell^{\mathrm{tri}}_{\mathrm{RA}}(\xi,y^a) = \Big[s(\mathbf{z}^a_\theta,\bar{\mathbf{z}}^-_k) - s(\mathbf{z}^a_\theta,\bar{\mathbf{z}}^+_k) + m\Big]_+ = \Big[m - \Delta_{\theta,k}(\xi,y^a)\Big]_+, \tag{9}$$

where $[u]_+ = \max(0,u)$ and $m > 0$ is the margin. Thus, $\phi_{\mathrm{tri}}(\Delta) = [m-\Delta]_+$.

3.2.3 Preference-consistent Target Gating

To avoid aligning the on-policy anchor with relational targets that are inconsistent with the current policy, we retain a preference pair for representation alignment only when its current DPO implicit reward preserves the annotated preference. Specifically, we define a preference-consistency gate as,

$$g_k(\xi) = \mathbb{1}\left[r^{\mathrm{DPO}}_{\theta_k}(x,y^+) > r^{\mathrm{DPO}}_{\theta_k}(x,y^-)\right], \tag{10}$$

where r^{DPO}_θ denotes the implicit reward associated with the underlying DPO objective. The detached gate affects only representation alignment. The gated representation-alignment objective is

$$\mathcal{L}^{\phi}_{\mathrm{RA}}(\theta;\theta_k) = \frac{\mathbb{E}_{(\xi,y^a)\sim Q_k}\left[g_k(\xi)\,\phi(\Delta_{\theta,k}(\xi,y^a))\right]}{\mathbb{E}_{\xi\sim\mathcal{D}}[g_k(\xi)]}, \tag{11}$$

where $\phi \geq 0$ is a non-increasing margin loss with $\phi(0) > 0$.

3.2.4 Alignment Guarantee

The RA objective is designed to enforce the desired representation ordering, i.e., making the on-policy anchor closer to the chosen target than to the rejected target. Minimizing the gated representation-alignment objective directly controls violations of this ordering on the preference-consistent subset. A representation-ordering violation occurs when $\Delta_{\theta,k}(\xi,y^a) \leq 0$. We thus define the preference-consistent representation alignment risk as

$$\mathcal{R}^{g}_{\mathrm{align}}(\theta;\theta_k) = \Pr_{(\xi,y^a)\sim Q_k}\left[\Delta_{\theta,k}(\xi,y^a) \leq 0 \mid g_k(\xi)=1\right]. \tag{12}$$

Theorem 1 (Preference-consistent alignment-risk bound)**.** *Assume* $\Pr_{\xi\sim\mathcal{D}}[g_k(\xi)=1] > 0$*. For any nonincreasing margin loss* $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ *with* $\phi(0) > 0$*,*

$$\Pr_{Q_k}\left[\Delta_{\theta,k} \leq 0 \mid g_k(\xi)=1\right] \leq \frac{\mathcal{L}^{\phi}_{\mathrm{RA}}(\theta;\theta_k)}{\phi(0)}. \tag{13}$$

Therefore, minimizing the representation objective reduces an upper bound on representation-ordering violations over the preference-consistent subset selected by g_k. The proof is provided in Appendix B.2.

Algorithm 1 On-Policy Representation Alignment (OPRA)

Require: $\mathcal{D}$, π_{ref}, θ_0, maximum anchor length K, coefficients β, λ, α, temperature τ or margin m

for $k = 0, \ldots, T-1$ **do**
  Sample $\{\xi_i = (x_i, y_i^+, y_i^-)\}_{i=1}^B \sim \mathcal{D}$ and $y_i^a \sim \widetilde{\pi}_{\theta_k}^{(\le K)}(\cdot|x_i)$ without gradient
  Forward on $(x_i, y_i^\pm)$ to compute $\bar{r}_i^\pm$ and detached targets $\bar{\mathbf{z}}_i^\pm$
  Construct $g_i = \mathbb{1}[\bar{r}_i^+ > \bar{r}_i^-]$
  Forward on (x_i, y_i^a) to compute $\mathbf{z}_i^a$ and $\Delta_i = s(\mathbf{z}_i^a, \bar{\mathbf{z}}_i^+) - s(\mathbf{z}_i^a, \bar{\mathbf{z}}_i^-)$
  Compute $\widehat{\mathcal{L}}_{\text{RA}} = \frac{\sum_i g_i \phi(\Delta_i)}{\sum_i g_i}$ if $\sum_i g_i > 0$, and set $\widehat{\mathcal{L}}_{\text{RA}} = 0$ otherwise
  Compute $\widehat{\mathcal{L}}_{\text{OPRA}} = \widehat{\mathcal{L}}_{\text{DPO}} + \lambda \widehat{\mathcal{L}}_{\text{RA}} + \alpha \widehat{\mathcal{L}}_{\text{NLL}}$
  Update $\theta_{k+1} \leftarrow \theta_k - \eta \nabla_\theta \widehat{\mathcal{L}}_{\text{OPRA}}$
end for

3.3 Overall OPRA Objective

To avoid the length sensitivity, we follow Meng et al. (2024) and normalized DPO objective to provide a more stable reward space for alignment Besides, we also follow Xu et al. (2024a); Meng et al. (2024) and employ the negative log-likelihood term on chosen responses to strength the generations. OPRA can be combined with either the standard or length-normalized DPO objective. With the length-normalized implicit reward $\bar{r}_\theta(x, y) = \frac{\beta}{|y|} \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$, the preference ranking and chosen-response objectives are

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\mathcal{D}} \left[\log \sigma \left(\bar{r}_\theta(x, y^+) - \bar{r}_\theta(x, y^-) \right) \right], \qquad \mathcal{L}_{\text{NLL}} = -\mathbb{E}_{\mathcal{D}} \left[\frac{1}{|y^+|} \log \pi_\theta(y^+|x) \right]. \quad (14)$$

The final OPRA objective is

$$\mathcal{L}_{\text{OPRA}}(\theta; \theta_k) = \mathcal{L}_{\text{DPO}}(\theta) + \lambda \mathcal{L}_{\text{RA}}(\theta; \theta_k) + \alpha \mathcal{L}_{\text{NLL}}(\theta), \quad (15)$$

where $\lambda, \alpha \ge 0$ and $\mathcal{L}_{\text{RA}}$ is the normalized gated representation objective in Eq. 11. Algorithm 1 summarizes the training procedure.

4 Experiments

4.1 Experimental Setups

Evaluation Benchmarks. We experiment on ten benchmarks, including seven knowledge and reasoning benchmarks, **MMLU** (Hendrycks et al., 2021), **ARC-Challenge (ARC-C)** and **ARC-Easy (ARC-E)** (Clark et al., 2018), **CommonsenseQA (CSQA)** (Talmor et al., 2019), **HellaSwag (HS)** (Zellers et al., 2019), **TruthfulQA (TQA)** (Lin et al., 2022), and **Winograd Schema Challenge (WS)** (Levesque et al., 2012), as well as three generation and instruction-following benchmarks **AlpacaEval2.0** (Li et al., 2023), **MT-Bench** (Zheng et al., 2023), and **IFEval** (Zhou et al., 2023). Please see Appendix C.1 for details of benchmarks and metrics.

Comparison Baselines. We compare OPRA with several preference optimization methods, including supervised-finetuning tuning **SFT**, **DPO** (Rafailov et al., 2023), **SimPO** (Meng et al., 2024), **ORPO** (Hong et al., 2024), **AlphaDPO** (Wu et al., 2025), **TIDPO** (Yang et al., 2026a), and **HyPO** (Yuan et al., 2026). Details can be found in Appendix C.2.

Models. We experiment on various SFT LLMs including **Qwen2.5-7B-Instruct** (Yang et al., 2024), **Gemma2-9B-it** (Team, 2024), **Llama3-8B-Instruct** (AI@Meta, 2024), **Mistral2-7B** (Jiang et al., 2023) and **Qwen3** (Yang et al., 2025). See Appendix C.3 for details.

Preference Data Settings. For the main experiments, we use two annotated preference datasets: **UltraFeedback** (Cui et al., 2024)[1], and **ORPO** dataset[2] (Hong et al., 2024). Each example contains a prompt and an offline chosen-rejected response pair with annotated preference reward scores provided by Djuhera et al. (2026). By default, we apply reward-consistency filtering using the provided

[1] https://huggingface.co/datasets/HuggingFaceH4/ultrafeedback_binarized

[2] https://huggingface.co/datasets/mlabonne/orpo-dpo-mix-40k

Table 1: Results (%) on seven knowledge and understanding benchmarks using UltraFeedback.

Method	MMLU	ARC-C	ARC-E	CSQA	HS	TQA	WS	Avg.	Δ_p
backbone: Qwen-2.5-7B-Instruct									
SFT	71.68	55.12	81.36	82.80	80.55	64.62	71.11	72.46	0.00
DPO	71.65	58.96	82.07	82.72	82.08	67.90	71.03	73.77	+2.08
DPO w/ L-norm	70.98	59.47	61.62	81.98	78.01	63.78	67.88	69.10	-3.91
DPO w/ NLL	71.69	59.30	83.25	83.05	82.05	67.78	71.19	74.04	+2.44
SimPO	70.88	68.69	84.97	82.15	82.91	72.28	74.35	76.60	+6.64
ORPO	70.93	59.56	84.01	83.70	79.72	63.81	73.88	73.66	+1.85
AlphaDPO	70.76	65.61	82.49	80.84	82.47	69.92	74.11	75.17	+4.51
TIDPO	71.67	55.63	81.19	82.56	80.79	65.47	70.17	72.50	+0.10
HyPO	71.40	61.18	80.35	82.56	82.22	72.63	71.43	74.54	+3.43
OPRA	71.70	71.50	86.99	83.37	82.68	73.82	71.35	**77.35**	**+7.80**
backbone: Gemma-2-9B-it									
SFT	71.84	65.02	84.22	80.18	80.04	60.19	76.32	73.97	0.00
DPO	72.38	69.03	86.41	80.59	83.38	67.36	77.03	76.60	+3.86
DPO w/ L-norm	71.13	65.96	69.40	78.05	69.43	69.25	70.64	70.55	-3.64
DPO w/ NLL	71.86	70.05	88.89	79.93	83.51	66.54	78.06	76.98	+4.31
SimPO	69.70	69.71	78.20	73.79	70.92	64.69	81.06	72.58	-1.23
ORPO	70.93	63.23	82.91	79.44	80.39	65.56	78.45	74.42	+0.81
AlphaDPO	70.28	67.49	75.63	70.84	68.72	63.32	80.90	71.03	-3.31
TIDPO	71.46	65.78	84.51	80.26	81.05	60.34	75.14	74.08	+0.15
HyPO	71.82	76.19	87.21	78.30	85.93	75.51	81.69	79.52	+8.32
OPRA	72.11	76.37	87.75	78.46	86.04	75.77	80.51	**79.57**	**+8.39**

FsfairX-RM scores and retain pairs satisfying $R_{\mathrm{RM}}(x, y^+) > R_{\mathrm{RM}}(x, y^-)$. All methods under the same setting are trained on the same filtered data. We further consider several alternative preference data settings including response generation, different annotators, and the effect of filtering for data quality. Details are reported in Appendix C.4.

Details of training settings are listed in Appendix C.5.

4.2 Main Results

Performance on Knowledge and Reasoning Tasks. Table 1 shows overall results of baselines and OPRA on seven knowledge and understanding benchmarks. OPRA consistently achieves the best average performance across both backbones. Compared with standard SFT, it improves the average score/Δ_p by **+4.9%/+7.8%** under Qwen2.5-7B-Instruct and **+5.6%/+8.3%** on Gemma2-9B-IT. The results demonstrate its superiority for preference optimization and generalization to LLM families.

Performance on Generation and Instruction-following Tasks. Table 3 shows results on three benchmarks for generation and instruction following. OPRA achieves superior or competitive overall performance compared to baselines across generation tasks. With Qwen2.5-7B-Instruct, OPRA outperforms DPO by **+1.8/0.8**% on IFEval in prompt/instance-level strict accuracy, **+0.3** on MT-Bench, and **+1.4/1.0**% in LR/WR on AlpacaEval2.0.

Performance against Different Training Preference Datasets. To evaluate robustness to different preference-data distributions, we further evaluate methods using ORPO dataset (Djuhera et al., 2026). As shown in Table 2 (see Appendix D.1 for detailed results for each benchmark), OPRA maintains consistent improvements across both datasets, indicating good generalization across different preference data sources.

Table 2: Results (%) on 7 knowledge and understanding benchmarks using ORPO dataset, with Qwen-2.5-7B-Instruct and Gemma-2-9B-it.

Method	Qwen-2.5		Gemma-2	
	Avg.	Δp	Avg.	Δp
SFT	72.46	0.00	73.97	0.00
DPO	72.90	+0.68	75.82	+2.71
HyPO	71.05	-1.78	75.68	+3.18
SimPO	73.70	+2.25	73.41	-0.02
AlphaDPO	72.16	+0.17	74.29	+1.13
ORPO	73.76	+1.91	74.25	+0.66
OPRA	**74.54**	**+3.26**	**76.50**	**+3.68**

We additionally analyze the **computational cost** during training in Appendix D.2. Results of efficiency show that OPRA improves alignment performance with acceptable training cost.

Table 3: Results on three generation benchmarks using UltraFeedback with Qwen-2.5-7B-Instruct.

Method	IFEval		MT-Bench			AlpacaEval2.0	
	Acc$_{\text{p}}$ (%)	**Acc**$_{\text{inst.}}$ (%)	Turn 1	Turn 2	**Overall**	**LC** (%)	**WR** (%)
DPO	70.24	78.90	7.74	6.60	7.22	50.72	50.93
SimPO	41.04	54.68	4.96	4.50	4.74	12.32	12.72
ORPO	64.88	73.62	7.40	**6.94**	7.18	50.15	44.71
AlphaDPO	43.07	56.59	5.02	5.13	5.07	15.35	15.82
TIDPO	70.98	79.02	8.00	6.83	7.44	49.64	49.07
HyPO	58.60	69.54	7.61	6.91	7.29	**52.26**	**51.93**
OPRA (Ours)	**72.09**	**79.74**	**8.04**	6.85	**7.48**	52.14	**51.93**

Table 5: Comparison results (%) against different representation-level losses on seven benchmarks.

Method	MMLU	ARC-C	ARC-E	CSQA	HS	TQA	WS	Avg.	Δ_p
backbone: Qwen-2.5-7B-Instruct									
OPRA	71.70	71.50	86.99	83.37	82.68	73.82	71.35	**77.35**	**+7.80**
w/o Gating	71.30	70.90	87.12	82.56	83.53	72.61	70.24	76.89	+7.11
w/o $\mathcal{L}_{\text{RA}}$	71.64	67.41	85.27	82.80	82.58	71.72	70.96	76.05	+5.76
w/ Positive	71.12	68.17	86.91	82.23	82.71	73.53	71.03	76.53	+6.48
w/ Soft	71.05	68.26	85.48	82.15	83.38	73.54	70.88	76.39	+6.32
w/ Hinge Triple	71.29	69.54	87.12	83.13	82.77	74.34	71.43	77.09	+7.33
backbone: Gemma-2-9B-it									
OPRA	72.11	76.37	87.75	78.46	86.04	75.77	80.51	**79.57**	**+8.39**
w/o Gating	70.65	77.39	90.15	77.81	86.41	71.37	76.80	78.65	+6.94
w/o $\mathcal{L}_{\text{RA}}$	70.50	72.18	87.16	79.20	84.09	74.39	77.11	77.80	+5.87
w/ Positive	70.99	72.35	84.43	75.92	84.56	77.11	76.64	77.43	+5.60
w/ Soft	70.69	74.49	87.42	77.07	85.37	74.99	76.09	78.02	+6.26
w/ Hinge Triple	71.88	74.74	86.03	76.25	84.53	74.99	78.69	78.16	+6.51

4.3 Analysis of On-Policy Representation Alignment

Comparison with Existing On-policy Alignment Strategy at Reward Level. To evaluate the advantage of representation-level on-policy supervision, we replace RA with the reward-level strategy in TIDPO (Yang et al., 2026a), keeping all other settings unchanged. From Table 4, OPRA achieves better overall performance, indicating that representation-level on-policy supervision provides richer and more effective preference signals than scalar reward-level alignment.

Table 4: Comparison with existing reward-level on-policy strategy (Yang et al., 2026a) on seven benchmarks.

Method	Avg.	Δp
OPRA	77.35	+7.80
w/ Triple (reward-level)	72.82	+0.62

Evaluation with Different Representation Losses. Table 5 compares several implementations of the representation alignment objective. First, we respectively remove the entire RA and gating. Both ablations lead to clear performance degradation, confirming the importance of RA and gating strategy. Then, we replace the default NCE to implement L_{RA} with three alternatives: (i) *Positive* only pulls the on-policy representation toward the chosen response, $\mathcal{L}_{\text{pos}} = 1 - s^+$, (ii) *Soft* adopts a smooth margin-ranking objective, $\mathcal{L}_{\text{soft}} = \log(1 + \exp((s^- - s^+ + m)/\tau))$, and (iii) *Hinge Triple* uses its hard-margin counterpart, $\mathcal{L}_{\text{tri}} = \max(0, s^- - s^+ + m)$. From results, positive-only alignment is weaker, showing that simply matching the preferred representation is insufficient. Both soft and triple-based losses improve over positive-only alignment, further supporting the im-

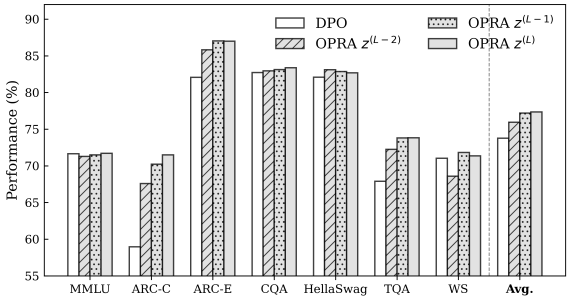

Figure 3: Results against hidden layers on seven knowledge and understanding benchmarks.

Table 7: Average results on seven benchmarks for methods trained under different preference data settings with Qwen2.5-7B-Instruct. * indicates the original dataset without quality filtering.

Method	UltraFeedback* Original annotation		UltraFeedback Original annotation		ORPO* Original annotation		ORPO Original annotation		UltraFeedback *GPT-4o*-annotated	
	Avg.	**Δp**	**Avg.**	**Δp**	**Avg.**	**Δp**	**Avg.**	**Δp**	**Avg.**	**Δp**
DPO	73.34	+1.42	73.77	+2.08	72.94	+0.79	72.90	+0.68	73.77	+2.08
SimPO	72.33	+0.48	76.60	+6.64	73.78	+2.36	73.70	+2.25	75.86	+5.59
HyPO	75.39	+4.62	74.54	+3.43	71.68	-0.90	71.05	-1.78	74.95	+3.90
AlphaDPO	71.88	-0.19	75.17	+4.51	72.41	+0.31	72.16	+0.17	75.07	+4.47
ORPO	72.99	+0.84	73.66	+1.85	72.85	+0.60	73.76	+1.91	72.93	+0.67
OPRA	**77.14**	**+7.27**	**77.35**	**+7.80**	**75.81**	**+5.33**	**74.54**	**+3.26**	**77.31**	**+7.58**

portance of relative preference modeling. Among them, the default NCE-based RA achieves better performance across backbones. These results demonstrate contrastive modeling of preference relations in representation space over simple matching or fixed-margin ranking.

Comparison of Representation Alignment at Different Hidden State Layers. We evaluate which layer provides the most effective representation for preference alignment. We apply the RA objective to hidden states from different layers. As shown in Figure 3, with Qwen2.5-7B-Instruct, performance generally improves as alignment moves to deeper hidden layers, with the final hidden state achieving the best overall results. This suggests that the last hidden layer preserves more informative preference information before output projection.

4.4 Evaluation across Preference Data Settings

Response Generation: Offline vs. On-policy. To examine the performance when the offline-to-on-policy mismatch is reduced from the data side, we compare offline preference data with a more on-policy setting following Meng et al. (2024). From Table 6, baseline methods generally benefit from such data, confirming that data matching can partially alleviate the mismatch. OPRA achieves the strongest performance under both settings, indicating that representation-level alignment provides a more effective and robust way to mitigate the mismatch with sufficient preference modeling.

Table 6: Results on seven benchmarks using *offline* and *on-policy* training data.

Method	Llama-3-8B-Instruct On-policy Data		Llama-3-8B-Instruct Offline Data	
	Avg.	**Δp**	**Avg.**	**Δp**
SFT	69.72	0.00	69.72	0.00
DPO	71.26	+2.77	70.50	+1.52
SimPO	72.97	+5.60	68.04	-1.48
OPRA	**73.41**	**+6.48**	**74.74**	**+8.28**

Different Preference Annotators. To study sensitivity to the source of preference supervision, we compare the original annotations in UltraFeedback with its GPT-4o re-annotated version (Peng et al., 2026). Results in Table 7 shows that OPRA remains effective under both annotation settings, demonstrating robustness to different preference judgments.

Reward Filtering: Unfiltered vs. Filtered. Table 7 compares the original preference data with its reward-filtered version to assess sensitivity to data filtering. OPRA remains consistently strong under both datasets, showing that OPRA is relatively insensitive to reward filtering.

Additionally, we evaluate the performance under **limited preference supervision** in Appendix D.3, showing that OPRA remains effective even when preference data are limited. We further study **different LLM scales**, **model families**, and **full fine-tuning paradigm** in Appendix D.4, with results consistently demonstrating the broad generalizability of OPRA across these settings.

5 Conclusion

This paper proposes OPRA framework to address the mismatch between offline preference supervision and current-policy generation from the representation perspective. It introduces policy-generated prefixes as on-policy anchors and directly aligns their representations with offline preferred and rejected responses. Experiments on ten various benchmarks show consistent improvements across multiple LLM backbones. Further analysis demonstrate the effectiveness of contrastive modeling of preference relations in representation space, and broad generalizability of OPRA. We discuss the limitations and future works in Appendix E.

AI USE STATEMENT

In this work, generative AI tools were used to aid and polish the writing, including grammar correction and language refinement. All AI-assisted text was reviewed and revised by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed OPRA in Section 3. Experimental settings, including evaluation datasets, metrics, training data, LLM backbones, hyperparameters required to reproduce our results, are reported in Section 4.1 and Appendix C. We will release the source code to further facilitate reproducibility.

REFERENCES

AI@Meta. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Zhaohan Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. Implicit preference optimization. *arXiv preprint arXiv:2405.14734*, 2024.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Peter Clark et al. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: boosting language models with scaled ai feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 9722–9744, 2024.

Aladin Djuhera, Farhan Ahmed, Swanand Kadhe, Syed Zawad, Heiko Ludwig, and Holger Boche. When data is the algorithm: A systematic study and curation of preference optimization datasets. In *International Conference on Learning Representations*, volume 2026, pp. 150085–150130, 2026.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *International Conference on Machine Learning*, pp. 12634–12651. PMLR, 2024.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Dan Hendrycks et al. Measuring massive multitask language understanding. *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. *KR*, 2012(13th):3, 2012.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics*, pp. 3214–3252, 2022.

Aiwei Liu, Haoping Bai, Zhiyun Lu, Yanchao Sun, Xiang Kong, Xiaoming Wang, Jiulong Shan, Albin Madappally Jose, Xiaojiang Liu, Lijie Wen, et al. Tis-dpo: Token-level importance sampling for direct preference optimization with estimated weights. In *International Conference on Learning Representations*, volume 2025, pp. 51339–51368, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Haozheng Luo, Yimin Wang, Jiahao Yu, Binghui Wang, and Yan Chen. Contrastive reasoning alignment: Reinforcement learning from hidden representations. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=56cQPBIg4R.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: simple preference optimization with a reference-free reward. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pp. 124198–124235, 2024.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017, 2024.

S. Peng, W. Wang, Z. Tian, S. Yang, X. Wu, H. Xu, C. Zhang, T. Isobe, B. Hu, and M. Zhang. Uni-dpo: A unified paradigm for dynamic preference optimization of llms. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=G7DBGlgjjp.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in llm reinforcement learning. *arXiv preprint arXiv:2602.04879*, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–823, June 2015.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4149–4158, 2019.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*, 2024.

Gemma Team. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118, 2024.

A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, 2024.

Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Alphadpo: Adaptive reward margin for direct preference optimization. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. ReFT: Representation finetuning for language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=fykjplMc0V.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54715–54754. PMLR, 21–27 Jul 2024.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 55204–55224, 2024a.

J. Xu, A. Lee, S. Sukhbaatar, and J. Weston. Swepo: Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682*, 2024b.

Zhongwen Xu and Zihan Ding. Single-stream policy optimization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=b61UW62K7W.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng

Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Ning Yang, Hai Lin, Yibo Liu, Baoliang Tian, Guoqing Liu, and Haijun Zhang. Token-importance guided direct preference optimization. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=cMEnMVvMw9.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Haobo Wang, Mingxuan Xia, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, Junbo Zhao, et al. Oprd: On-policy representation distillation. *arXiv preprint arXiv:2606.06021*, 2026b.

Suqin Yuan, Xingrui Yu, Jiyang Zheng, Lei Feng, Dadong Wang, Ivor Tsang, and Tongliang Liu. Mitigating mismatch within reference-based preference optimization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=k79Un1LSXy.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. Self-rewarding language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57905–57923. PMLR, 21–27 Jul 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 58348–58365, 2024.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. Wpo: Enhancing rlhf with weighted preference optimization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8328–8340, 2024.

Mingkang Zhu, Xi Chen, Zhongdao Wang, Bei Yu, Hengshuang Zhao, and Jiaya Jia. Tgdpo: Harnessing token-level reward guidance for enhancing direct preference optimization. In *International Conference on Machine Learning*, pp. 79727–79745. PMLR, 2025a.

Wenqiao Zhu, Ji Liu, Rongjunchen Zhang, Haipang Wu, and Yulun Zhang. Carft: boosting llm reasoning via contrastive learning with annotated chain-of-thought-based reinforced fine-tuning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5933–5948, 2025b.

APPENDIX

In the appendix, we will provide: (i) related works, (ii) theory proofs, (iii) detailed experimental settings, and (iv) additional results.

A RELATED WORK

A.1 PREFERENCE OPTIMIZATION

Aligning large language models (LLMs) with human preferences has become a standard paradigm in post-training. Existing alignment methods can generally be divided into *explicit reward optimization* and *implicit preference optimization*.

Explicit Reward Optimization Explicit reward optimization methods, such as reinforcement learning from human feedback (RLHF), learn reward functions and optimize policies through reinforcement learning. Representative approaches including PPO (Schulman et al., 2017), GRPO (Shao et al., 2024), and related variants (Zheng et al., 2025; Gao et al., 2025; Xu & Ding, 2026; Qi et al., 2026) achieve strong alignment performance but often require additional reward modeling and iterative policy updates.

Implicit Reward Optimization Direct Preference Optimization (DPO) formulates LLM alignment through policy-reference log-probability ratios, avoiding explicit reward-model training and policy-gradient optimization (Rafailov et al., 2023). Subsequent implicit optimization methods extend this formulation from different perspectives. SimPO (Meng et al., 2024) removes the dependency on reference models through normalized likelihood optimization. Some methods exploit richer supervision structures (Xu et al., 2024b;a). IPO (Azar et al., 2024) revisits preference learning from an implicit reward perspective, while KTO (Ethayarajh et al., 2024) extends optimization beyond pairwise preference supervision. R-DPO (Park et al., 2024) introduces regularization to mitigate undesired length bias, and Uni-DPO (Peng et al., 2026) improves preference utilization through adaptive quality-aware weighting. Recent works improve finer-grained preference supervision beyond sequence-level objectives (Zeng et al., 2024; Zhu et al., 2025a; Liu et al., 2025; Yang et al., 2026a).

Offline-to-on-policy Mismatch in DPO-style Paradigm However, strong preference classification on offline pairs does not necessarily guarantee the high-quality policy generations (Tang et al., 2024). To mitigate this mismatch, existing works mainly modify either the training data distribution or the supervision applied to generated responses. For example, some studies improves data collection from the initial policy. Guo et al. (2024); Rosset et al. (2024); Xiong et al. (2024); Yuan et al. (2024) collect policy-generated responses and obtain preference feedback iteratively. Meng et al. (2024) construct preference datasets from the policy initialization. Zhou et al. (2024) re-weight offline pairs according to current-policy probabilities. For objectives, (Yang et al., 2026a) develop a token-level log-ratios objective on an on-policy anchor to chosen and rejected responses using triplet constraints These approaches bring supervision closer to policy behavior while retaining reward-level optimization targets.

Different previous works, this paper explores a different representation perspective and develops on-policy representation alignment to explicitly model preference relations in representation space.

A.2 REPRESENTATION LEARNING AND ALIGNMENT

Directly modeling LLM representations has been explored for task adaptation, safety, and reasoning. Wu et al. (2024) learn task-specific interventions on hidden representations (Zhu et al., 2025b) model representations of chain-of-thought for boosting RL-based fine-tuning. More recently, contrastive representation learning has also been integrated with reinforcement learning. (Luo et al., 2026) integrate contrastive representation learning with reinforcement learning to learn class-defined latent structure for safety alignment. (Yang et al., 2026b) utilize knowledge distillation to encourage representation matching between the output space to the hidden-state space. Unlike these works, this paper focuses on implicit preference optimization and explore on-policy representation learning to mitigate the mismatch between on-policy generation and offline preference supervision in DPO-style paradigm.

B Theory Proof

B.1 Proof of Proposition 1

We show that whenever a representation relation contains a first-order variation not captured by the output statistics, there exists a parameter direction that changes the representation relation while leaving the output statistics unchanged to first order.

Proof. Let $J_M = \partial \mathbf{M}_\theta / \partial \theta$. Since $\mathrm{row}(J_M)^\perp = \ker(J_M)$, decompose

$$\nabla_\theta g_\theta = J_M^\top \mathbf{a} + \mathbf{w}, \qquad \mathbf{w} \in \ker(J_M).$$

The assumption $\nabla_\theta g_\theta \notin \mathrm{row}(J_M)$ implies $\mathbf{w} \neq 0$. Choose $\mathbf{v} = \mathbf{w}$. By a first-order Taylor expansion,

$$\mathbf{M}_{\theta+\epsilon\mathbf{v}} = \mathbf{M}_\theta + \epsilon J_M \mathbf{v} + O(\epsilon^2) = \mathbf{M}_\theta + O(\epsilon^2), \tag{16}$$

$$g_{\theta+\epsilon\mathbf{v}} = g_\theta + \epsilon \nabla_\theta g_\theta^\top \mathbf{v} + O(\epsilon^2). \tag{17}$$

Moreover,

$$\nabla_\theta g_\theta^\top \mathbf{v} = (J_M^\top \mathbf{a} + \mathbf{w})^\top \mathbf{w} = \|\mathbf{w}\|_2^2 > 0,$$

because $J_M \mathbf{w} = 0$. Thus, the output statistics are unchanged to first order while the representation relation changes to first order. □

B.2 Proof of Theorem 1

We prove Theorem 1 by upper-bounding the binary representation-ordering violation with the non-increasing margin loss and then taking expectations over the gated on-policy samples.

Proof. Since ϕ is non-increasing and $\phi(0) > 0$, for any $\Delta \in \mathbb{R}$,

$$\mathbb{1}\{\Delta \leq 0\} \leq \frac{\phi(\Delta)}{\phi(0)}. \tag{18}$$

Indeed, when $\Delta \leq 0$, we have $\phi(\Delta) \geq \phi(0)$; otherwise the left-hand side is zero.

Applying this inequality to $\Delta_{\theta,k}(\xi, y^a)$, multiplying by $g_k(\xi)$, and taking expectations over $(\xi, y^a) \sim Q_k$ gives

$$\mathbb{E}_{Q_k}\left[g_k(\xi)\mathbb{1}\{\Delta_{\theta,k}(\xi, y^a) \leq 0\}\right] \leq \frac{\mathbb{E}_{Q_k}\left[g_k(\xi)\phi(\Delta_{\theta,k}(\xi, y^a))\right]}{\phi(0)}. \tag{19}$$

Since the marginal of Q_k over ξ is $\mathcal{D}$, $\mathbb{E}_{Q_k}[g_k(\xi)] = \mathbb{E}_{\mathcal{D}}[g_k(\xi)]$. Dividing by this positive quantity yields

$$\Pr_{Q_k}\left[\Delta_{\theta,k}(\xi, y^a) \leq 0 | g_k(\xi) = 1\right] \leq \frac{\mathcal{L}_{\mathrm{RA}}^{\phi}(\theta; \theta_k)}{\phi(0)}, \tag{20}$$

which is exactly $\mathcal{R}_{\mathrm{align}}^{g}(\theta; \theta_k) \leq \mathcal{L}_{\mathrm{RA}}^{\phi}(\theta; \theta_k)/\phi(0)$. □

C Detailed Experimental Setups

C.1 Details of Evaluation Benchmarks

The seven benchmarks covering reasoning and knowledge understanding includes: **MMLU** (Hendrycks et al., 2021), which evaluates broad multitask knowledge and reasoning across diverse academic domains; **ARC-Challenge (ARC-C)** and **ARC-Easy (ARC-E)** (Clark et al., 2018), which measure scientific question answering ability at different difficulty levels; **CommonsenseQA (CSQA)** (Talmor et al., 2019), which tests commonsense reasoning; **HellaSwag (HS)** (Zellers et al., 2019), which evaluates grounded commonsense inference and sentence completion; **TruthfulQA (TQA)** (Lin et al., 2022), which measures the ability to generate truthful responses and avoid common misconceptions; and **Winograd Schema Challenge (WS)** (Levesque

et al., 2012), which assesses pronoun resolution and commonsense reasoning. Following standard evaluation protocols, we use standard accuracy for MMLU, CommonsenseQA, and WS, normalized accuracy for ARC-C, ARC-E, and HellaSwag, and MC2 accuracy for TruthfulQA. To evaluate an overall performance, we compute the average scores (Avg.) of the seven benchmarks, and the average of the relative improvement (Δp) over SFT on each metric of each benchmark.

The three benchmarks are used to evaluate the generation quality and instruction-following ability, i.e., **IFEval** (Zhou et al., 2023) tests whether models accurately follow instructions. We report both prompt-level and instance-level strict accuracy, denoted as Acc_{p} and $\text{Acc}_{\text{inst.}}$. **MT-Bench** (Zheng et al., 2023) evaluates responses to multi-turn questions using an LLM judge. We use `gpt-5-nano` as the judge model and report the first-turn, second-turn, and overall scores. **AlpacaEval2.0** (Li et al., 2023) evaluates instruction-following quality through pairwise comparisons. We report win rate (WR) and length-controlled win rate (LC). We use the base model (i.e., Qwen2.5-7B-Instruct) as the reference model and `gpt-3.5-turbo-16k` as the evaluator.

C.2 Details of Baselines

We compare OPRA with several preference optimization methods. Supervised-finetuning tuning corresponding to the **SFT** model in our results, serves as the lower-bound reference. **DPO** (Rafailov et al., 2023) uses the log-likelihood ratio between the current policy and a reference model. **SimPO** (Meng et al., 2024) is a reference-free method and uses length-normalized average log-likelihood as the implicit reward. **TIDPO** (Yang et al., 2026a) estimates token-level importance using gradient attribution and a Gaussian prior, together with a on-policy triplet loss for token-level preference supervision. **ORPO** (Hong et al., 2024) jointly maximizes the likelihood of chosen responses and enforces preference separation via an odds-ratio objective. **AlphaDPO** (Wu et al., 2025) reparameterizes the reference distribution to construct sample-specific preference margins for more flexible preference optimization. **HyPO** (Yuan et al., 2026) reduces the influence of the reference model when it conflicts with the preference labels while retaining its regularization effect.

C.3 Details of Models

We experiment on various supervised fine-tuned (SFT) LLMs: **Qwen2.5-7B-Instruct** (Yang et al., 2024), **Gemma2-9B** (Team, 2024), **Llama3-8B-Instruct** (AI@Meta, 2024), **Mistral2-7B** (Jiang et al., 2023). Specifically, we utilize *meta-llama/Meta-Llama-3-8B-Instruct*, *Qwen/Qwen2.5-7B-Instruct*, *mistralai/Mistral-7B-Instruct-v0.3*, and *google/gemma-2-9b-it*). We use five pretrained instruction-tuned **Qwen3** (Yang et al., 2025) models of different scales: *Qwen/Qwen3-0.6B*, *Qwen/Qwen3-1.7B*, *Qwen/Qwen3-4B-Instruct-2507*, *Qwen/Qwen3-8B*, and *Qwen/Qwen3-14B*.

C.4 Details of Alternative Preference Data Settings

We consider the following alternative preference data settings to evaluate the effectiveness of OPRA. (i) To evaluate a setting closer to the on-policy distribution, we use the another version of UltraFeedback released with (Peng et al., 2026), where new responses are generated by Llama-3-8B-Instruct and annotated using ArmoRM (Wang et al., 2024) and GPT-4o (Hurst et al., 2024), making the response distribution closer to the policy being optimized. (ii) We use a GPT-4o re-annotated version of UltraFeedback released with Peng et al. (2026) to evaluate robustness to different annotations. (iii) To study the effect of preference-pair quality, we compare the default filtered versions of both datasets with their unfiltered versions.

C.5 Training Settings

All experiments are conducted on $4\times$ NVIDIA A800-80GB GPUs. For efficient fine-tuning, we adopt LoRA (Hu et al., 2022). We set the LoRA rank to $r = 64$, the scaling factor to 128, and the dropout rate to 0.05. The target modules include `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. All models are trained for one epoch. For Qwen2.5-7B-Instruct, we use a per-device batch size of 4 with 16 gradient accumulation steps. For Gemma2-9B-it, we use a per-device batch size of 2 with 32 gradient accumulation steps and the eager attention implementation. The maximum sequence length is 512, and the maximum prompt length is 256.

Table 8: Average scores on seven knowledge and understanding benchmarks. All methods are trained using ORPO preference dataset.

Method	MMLU	ARC-C	ARC-E	CSQA	HS	TQA	WS	Avg.	Δ_p
backbone: Qwen-2.5-7B-Instruct									
SFT	71.68	55.12	81.36	82.80	80.55	64.62	71.11	72.46	0.00
DPO	71.79	55.20	80.43	82.80	81.60	67.59	70.88	72.90	+0.68
SimPO	70.45	63.65	80.18	82.88	82.19	65.68	70.88	73.70	+2.25
ORPO	71.39	59.30	84.68	84.85	79.93	63.01	73.16	73.76	+1.91
AlphaDPO	69.35	61.77	75.59	80.43	81.30	65.16	71.51	72.16	+0.17
HyPO	71.59	51.79	68.60	82.72	81.45	69.53	71.67	71.05	-1.78
OPRA (ours)	72.06	61.01	83.67	83.78	80.22	68.72	72.30	**74.54**	**+3.26**
backbone: Gemma-2-9B-it									
SFT	71.84	65.02	84.22	80.18	80.04	60.19	76.32	73.97	0.00
DPO	72.87	66.38	84.47	80.18	83.59	65.65	77.58	75.82	+2.71
SimPO	68.97	67.06	78.16	72.24	76.09	69.64	81.69	73.41	-0.02
ORPO	72.23	63.82	82.58	78.46	79.36	65.85	77.43	74.25	+0.66
AlphaDPO	69.70	68.34	79.21	76.17	76.38	69.70	80.51	74.29	+1.13
HyPO	70.87	66.98	79.29	73.38	83.97	75.50	79.79	75.68	+3.18
OPRA (ours)	72.41	68.34	86.62	79.12	83.82	66.55	78.61	**76.50**	**+3.68**

We use AdamW (Loshchilov & Hutter, 2019) with a cosine learning-rate schedule, 10% warmup, and a learning rate of 1.0×10^{-4}.

OPRA Settings. The default OPRA uses the last-layer hidden states to construct response-level representations and cosine similarity for representation alignment. We use the NCE-based RA objective with temperature $\tau = 0.2$. The maximum on-policy anchor length is $K = 8$. Anchors are sampled from the current policy with temperature $T_{\text{gen}} = 0.2$, top-$k = 20$, and top-$p = 0.7$. We set $\beta = 0.1$. For the knowledge and reasoning benchmarks, we use length-normalized DPO. With Qwen2.5-7B-Instruct, we set $\lambda = 0.001$ and $\alpha = 0.01$; with Gemma2-9B-it, we set $\lambda = 0.2$ and $\alpha = 0.02$. For the generation and instruction-following benchmarks with Qwen2.5-7B-Instruct, we use DPO without length normalization and set $\lambda = 0.01$ and $\alpha = 0$.

Full-parameter Fine-tuning. For the full-parameter fine-tuning experiments, we disable LoRA and optimize all model parameters with DeepSpeed ZeRO-3. We use a learning rate of 1×10^{-5}, a per-device batch size of 2, and 32 gradient accumulation steps.

D Additional Experimental Results

D.1 Detailed Results of Methods with ORPO Dataset

Table 8 shows detailed results on each benchmark of methods trained with ORPO dataset.

D.2 Efficiency Analysis

We evaluate the additional computational cost introduced by on-policy representation alignment. We compare training time, and GPU memory usage during training. As shown in Table 9, OPRA introduces only moderate training overhead over DPO ($\times 1.32$ vs. $\times 1.00$), despite the additional representation alignment objective. Compared with baseline methods such as SimPO improving with on-policy data, they usually need an additional cost of constructing on-policy preference data when applicable. Compared with on-policy optimization method such as TIDPO, OPRA is considerably more efficient. Overall, OPRA provides substantial performance gains with acceptable computational overhead, achieving a favorable effectiveness-efficiency trade-off.

D.3 Generalization Evaluation under Limited Preference Data

To evaluate OPRA under limited preference supervision, we evaluate OPRA under limited training preference data. We randomly sample 20%, 40%, 60%, and 80% of the UltraFeedback training set,

Table 9: Results of efficiency analysis using Qwen2.5-7B-Instruct during training.

Method	Runtime	GPU (GB)
DPO/AlphaDPO/HyPO	×1.00	47.7
ORPO	×0.79	47.7
SimPO	×0.79	60.1
TIDPO	×4.66	54.3
OPRA	×1.32	58.3

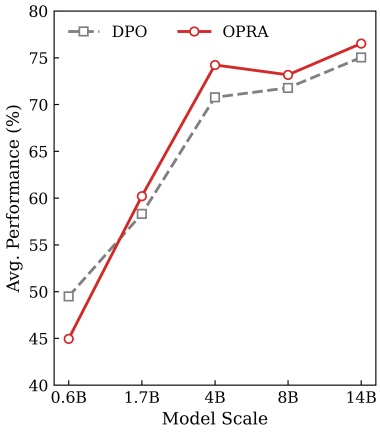

Figure 4: Performance against different LLM scales for Qwen3. All methods are trained with UltraFeedback.

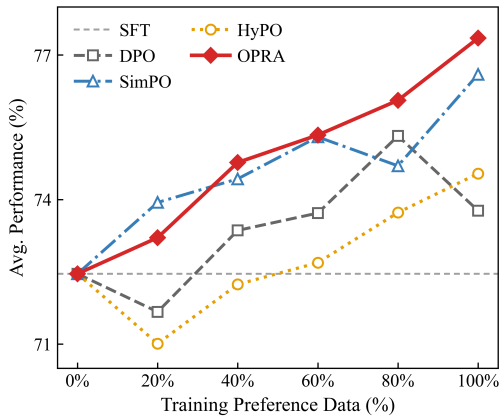

Figure 5: Results against different training ratios of preference data with Qwen2.5-7B-Instruct. All methods are trained with UltraFeedback.

using the same subsets for all compared methods while keeping all other settings unchanged. As shown in Figure 5, OPRA consistently outperforms DPO across all data ratios. The results show that OPRA remains effective even when preference data are limited.

D.4 Evaluation on Different LLM Families, LLM Scales, and Fine-tuning Paradigms

D.4.1 Evaluation against Different LLM Scales.

We evaluate the scalability of OPRA across different model size. We experiment with Qwen3 family, ranging from 0.6B to 14B parameters. As shown in Figure 4, OPRA consistently improves over DPO from 1.7B to 14B, demonstrating that OPRA generalizes well across abroad range of model scales and suggest promising scalability to larger models. OPRA underperforms DPO on the smallest 0.6B model, suggesting that limited model capacity may constrain the effectiveness of representation-level alignment.

D.4.2 Evaluation on Different LLM Families

We evaluate whether OPRA generalizes beyond the backbones used in the main experiments. Figure 6 (a) and (b) show results on Llama-3-8B-Instruct and Mistral-7B-Instruct under the default training setting. OPRA achieves broader performance coverage across the seven benchmarks and consistently outperforms DPO in terms of average performance, with gains of +2.48 and +3.06 points on the two backbones, respectively. This shows that OPRA generalizes across different LLM families.

D.4.3 Evaluation under Full-parameter Fine-Tuning

We additionally perform full-parameter fine-tuning to evaluate whether OPRA generalizes across different optimization paradigms. From Figure 6(c), OPRA maintains consistent improvements compared with baselines under full fine-tuning, demonstrating its generalization on both efficient and full-parameter fine-tunings.

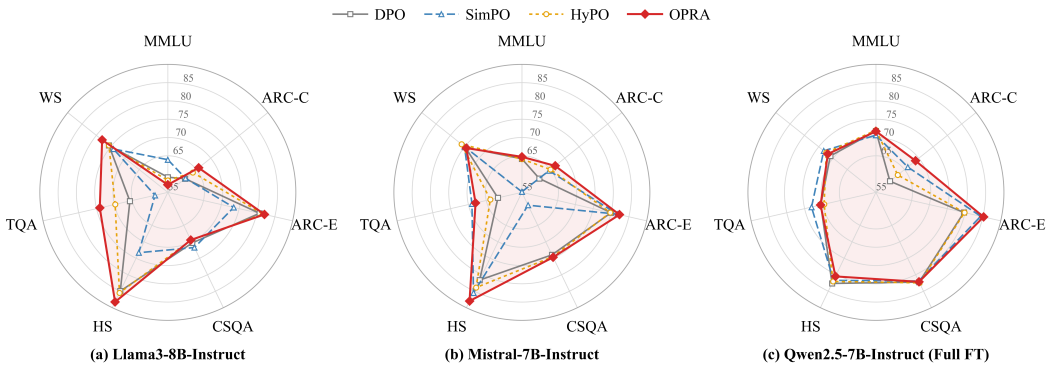

Figure 6: Results on seven knowledge and reasoning benchmarks (a-b) across other LLM families, and (c) using Qwen2.5-7B-Instruct under full-parameter fine-tuning. All methods are trained with UltraFeedback.

E Limitations and Future Directions

Like other on-policy methods, OPRA requires additional on-policy generation and representation computation during training, introducing moderate extra cost over standard offline preference optimization. Future work may explore more efficient representation losses. Additionally, due to budget and resource constraints, we use GPT-5 Nano and GPT-3.5 Turbo as the judge model instead of GPT-4o and primarily adopt LoRA for the main experiments. We evaluate models up to 14B parameters and consider three generation benchmarks. Extending OPRA to larger models such as 70B-scale LLMs, and broader generation benchmarks such as Arena-Hard, would further evaluate its generalizability.