MITIGATING MISMATCH WITHIN REFERENCE-BASED PREFERENCE OPTIMIZATION

Suqin Yuan[1] **Xingrui Yu**[2] **Jiyang Zheng**[1, 3]
Lei Feng[4]* **Dadong Wang**[3] **Ivor Tsang**[2, 5] **Tongliang Liu**[1]*
[1] Sydney AI Centre, The University of Sydney
[2] CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore
[3] Data61, CSIRO [4] Southeast University [5] CCDS, Nanyang Technological University
suqinyuan.cs@gmail.com, fenglei@seu.edu.cn, tongliang.liu@sydney.edu.au

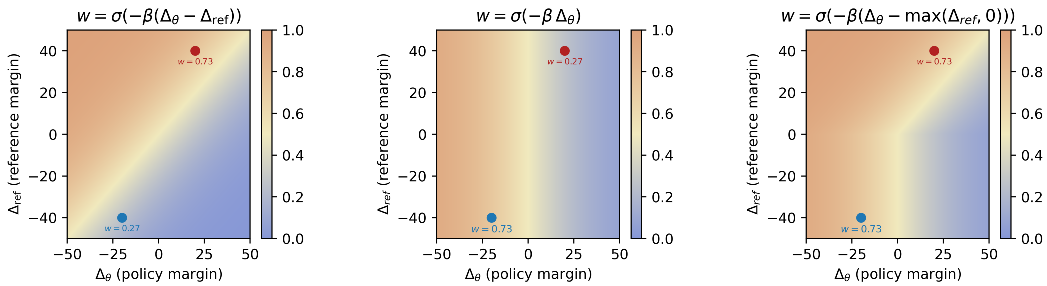
ABSTRACT

Direct Preference Optimization (DPO) has become the *de facto* standard for offline preference alignment of large language models, but its reliance on a reference policy introduces a critical tension. DPO weighs each update relative to a reference, which stabilizes the training by regularizing the updates within a trusted region. This reliance becomes problematic for pessimistic pairs, where the reference model prefers the rejected response. For these pairs, DPO prematurely attenuates the gradient as soon as the policy margin (Δ_θ) merely beats the reference margin (Δ_{ref}) even if the policy is still wrong ($\Delta_\theta < 0$). We name this failure *premature satisfaction*, which is a concrete form of the *training–inference mismatch*. Reference-free objectives remove this mismatch by optimizing the absolute margin, but at the cost of discarding the stabilizing signal of the reference. We mitigate this tension with *Hybrid-DPO (HyPO)*, a drop-in modification to DPO that applies reference *conditionally*: *HyPO* behaves exactly like DPO when the reference is optimistic or neutral, and it treats the reference as neutral when it is pessimistic by replacing $\Delta_\theta - \Delta_{\text{ref}}$ with $\Delta_\theta - \max\{0, \Delta_{\text{ref}}\}$. This one-line change strictly strengthens per-example learning signals on pessimistic pairs while preserving DPO's objective form and computational cost. By conditionally debiasing the pessimistic reference signal, *HyPO* mitigates *premature satisfaction*; empirically, across preference alignment, *HyPO* improves inference-aligned metrics and achieves higher pairwise win rates. Our results provide evidence that direct preference alignment could be enhanced by conditionally debiasing the reference signal, rather than discarding it.

1 INTRODUCTION

Aligning Large Language Models (LLMs) with human preferences and values is paramount for developing safe, helpful, and reliable AI systems (Askell et al., 2021). The pioneering Reinforcement Learning from Human Feedback (RLHF) paradigm (Christiano et al., 2017) achieved this by first training a reward model on human preference pairs (Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022; Bradley & Terry, 1952; Schulman et al., 2017; Bradley & Terry, 1952), and then using reinforcement learning (Schulman et al., 2017) to optimize the LLM for reward maximization, usually with a KL-divergence penalty to regularize deviation from a reference policy (Ouyang et al., 2022). Despite the widespread success, RLHF's multi-stage process suffers from complexity, instability, and high computational cost (Sharma et al., 2024; Gao et al., 2023). Addressing these challenges, Direct Preference Optimization (DPO) (Rafailov et al., 2023) emerged as a compelling alternative, offering a significant simplification by analytically deriving a direct policy loss. This loss allows for direct optimization on pairwise preference data (x, y_w, y_l), bypassing explicit reward modeling. Its mechanism implicitly enforces KL regularization by comparing the policy to a reference model (π_{ref}), making the loss dependent on the log-likelihood margin $\Delta_\theta - \Delta_{\text{ref}}$. DPO has gained widespread adoption due to its stability, efficiency, and strong empirical performance (Rafailov et al., 2023; D'Oosterlinck et al., 2024; Zhao et al., 2023; Yuan et al., 2023b; Rashidinejad & Tian, 2025).

*Corresponding authors.



(a) DPO (b) Reference-free Method (c) *HyPO* (Ours)

Figure 1: Per-example gradient weight heatmaps. The weight, plotted over the policy margin Δ_θ (x-axis) and the reference margin Δ_{ref} (y-axis), indicates training signal strength. DPO's reliance on the relative margin leads to *premature satisfaction*: on pessimistic examples (blue dot, $\Delta_{\text{ref}} < 0$), the signal is heavily attenuated even though the policy is still wrong in absolute terms. Reference-free methods provide a strong signal but discard the reference entirely. *HyPO* mitigates this by mirroring the reference-free behavior on pessimistic examples to ensure a strong signal, while reverting to DPO on optimistic examples (red dot) to maintain proximity to the reference policy.

While DPO is simple and effective, its reliance on a reference model introduces a tension that has led to two seemingly opposing lines of work. One argues that the reference should be removed due to a *training–inference mismatch* (Chen et al., 2024; Meng et al., 2024): DPO optimizes a reference-relative margin (e.g., $\Delta_\theta - \Delta_{\text{ref}}$), whereas inference depends on *absolute* log-likelihoods with no reference. Empirically, after DPO alignment, the agreement between the implicit-reward ordering and the likelihood ordering is only $\sim$50%, motivating reference-free objectives that directly optimize the inference metric and often outperform DPO (Meng et al., 2024). The other line takes the opposite view: the reference can be beneficial when strengthened or made adaptive. Updating π_{ref} alongside the policy helps mitigate over-optimization (Gorbatovski et al., 2025; He et al., 2026), and employing a stronger or better-matched reference improves stability and final performance, sometimes surpassing reference-free methods (Liu et al., 2024b; Azar et al., 2024; Pan et al., 2025). This leaves a practical dilemma: removing π_{ref} eliminates metric mismatch and can enhance performance, whereas retaining π_{ref} improves training stability and can also enhance performance.

To mitigate the tension between reference-based and reference-free direct alignment, we propose a unifying principle: the proximal stability offered by the reference model should be applied conditionally. We argue that the reference provides a constructive signal for stable learning only when it aligns with the preference data. When it is pessimistic ($\Delta_{\text{ref}} < 0$), its stabilizing pull becomes a misleading bias, making the learning objective unnecessarily difficult and causing the gradient to attenuate even when the absolute ordering (Δ_θ) is still incorrect. We term this effect *premature satisfaction*. Prior approaches either accept this bias (standard DPO; Rafailov et al., 2023) or discard potentially useful reference signal altogether (reference-free methods; Yuan et al., 2023b; Meng et al., 2024).

To realize this principle, we introduce *Hybrid-DPO* (*HyPO*), a minimalist and efficient modification to the DPO objective. *HyPO* behaves identically to DPO when the reference is helpful (optimistic or neutral) but treats a pessimistic reference as neutral ($\Delta_{\text{ref}} = 0$). This is achieved with a minor change to the DPO loss: the log-likelihood margin $\Delta_\theta - \Delta_{\text{ref}}$ becomes $\Delta_\theta - \max\{0, \Delta_{\text{ref}}\}$. As illustrated in Figure 1, standard DPO (a) prematurely attenuates the learning signal on pessimistic examples; typical reference-free method (b) discard the reference signal entirely; our proposed *HyPO* (c) uses the stabilizing reference where it helps and nullifies it where it is detrimental, thereby continuing to push the policy toward the correct absolute ordering.

We conduct a comprehensive empirical evaluation on preference datasets, aligning both base and instruction-tuned variants of two widely-used LLM family (Jiang et al., 2023; AI@Meta, 2024). Our proposed method, *HyPO*, demonstrates superior performance on general instruction-following benchmarks like AlpacaEval 2.0 (Li et al., 2023) and Arena-Hard-v0.1 (Li et al., 2024b). Notably, these results translate to a 41.2% average relative improvement over the original DPO. Further analyses and ablations corroborate the robustness of our approach, showing that *HyPO* maintains its advantage when scaling to larger models and transferring to different datasets, all while preserving performance on a wide array of downstream tasks. These findings empirically validate our central thesis: conditionally debiasing the pessimistic reference signal, rather than discarding it entirely, provides a more robust and effective path toward stable, high-performing direct preference alignment.

Our main contributions can be summarized as follows:

1. We identify and formalize *premature satisfaction*: for pessimistic pairs ($\Delta_{\mathrm{ref}} < 0$), the DPO weight attenuates once the *relative* margin turns positive even when the *absolute* margin remains wrong ($\Delta_\theta < 0$), revealing a tension between correctness and stability.

2. We propose *Hybrid-DPO* (*HyPO*), a minimalist modification to the DPO objective that clips the reference margin in the DPO loss, thereby preserving the DPO-style logistic loss and, on pessimistic pairs ($\Delta_{\mathrm{ref}} < 0$), reducing to an absolute margin update, with no additional computational cost.

3. Across large-scale preference alignment, *HyPO* improves inference-aligned metrics and raises pairwise win rates on general instruction-following benchmarks (e.g., AlpacaEval, Arena-Hard).

2 Related Work

We situate *HyPO* within the off-policy branch of RLHF: methods trained on fixed human preference pairs without on-policy rollouts. Accordingly, we do not compare to on-policy RLHF (Christiano et al., 2017; Shao et al., 2024), and instead focus on recent direct preference alignment (DPO-style) objectives and extensions, underscoring *HyPO*'s distinct features.

Manipulating the Reference Signal. DPO's reliance on π_{ref} has inspired a spectrum of solutions. One direction focuses on refining the reference: performance is sensitive to reference choice and strength (Liu et al., 2024b; Azar et al., 2024), leading to Trust-Region DPO (TR-DPO) (Gorbatovski et al., 2025), which dynamically updates π_{ref} during training to keep it aligned with the policy. RainbowPO (Zhao et al., 2025) mixes the standard reference with an idealized constant-margin policy. Pre-DPO (Pan et al., 2025) further refines the reference model's role as an implicit sample reweighter. At the other extreme, reference-free methods like RRHF (Yuan et al., 2023b), SLiC-HF (Zhao et al., 2023), CPO (Xu et al., 2024), ORPO (Hong et al., 2024a), and SimPO (Meng et al., 2024) eliminate the reference to directly optimize absolute log-likelihoods, addressing the *training–inference mismatch* at the cost of removing a stabilizing prior. The core problem of reference mismatch has also been identified in other modalities; for example, MaPO (She et al., 2024) removes the reference to improve multimodal alignment. *HyPO* offers a principled middle ground: it keeps the DPO-style logistic loss, but clips the reference margin, replacing Δ_{ref} with $\max\{0, \Delta_{\mathrm{ref}}\}$. This removes pessimistic anchoring while preserving proximality when the reference is non-pessimistic.

Explicit Reweighting and Robust Optimization. A parallel line of work makes the reference-induced sample weighting explicit to improve training dynamics and robustness. Several distributionally robust DPO variants (e.g., rDPO (Chowdhury et al., 2024), Dr.DPO (Wu et al., 2024a), WDPO (Xu et al., 2025)) reweight examples to minimize the impact of label noise (Yuan et al., 2023a; 2024; 2025b;a) or worst-case distributional shifts. Others use data-driven signals; for instance, β-DPO (Wu et al., 2024b) dynamically adjusts the temperature β based on data quality, while WPO (Zhou et al., 2024) reweights pairs to better simulate on-policy optimization from off-policy data. Some methods operate at a finer granularity, such as TDPO (Zeng et al., 2024) and TIS-DPO (Liu et al., 2024a), which apply importance sampling at the token level. While these methods all leverage reweighting, their criteria are typically noise, data quality, or on-policy correction. In contrast, *HyPO* gates the reference contribution by the reference model's own pessimism: when $\Delta_{\mathrm{ref}} \geq 0$ it behaves exactly like DPO, and when $\Delta_{\mathrm{ref}} < 0$ it reduces to an absolute margin update on those pairs.

Alternative Objective Formulations. Beyond manipulating the reference or loss weights, several methods alter the structure of the DPO objective. f-DPO generalizes DPO beyond reverse-KL to broader f-divergences (Wang et al., 2024a). R-DPO diagnoses and mitigates length biases within the DPO objective (Park et al., 2024), and AlphaDPO adaptively sets a data-dependent reward margin (Wu et al., 2025). Techniques such as DPO-Shift (Yang et al., 2025) and BDPO (Cho et al., 2025) modify the loss function to bound or rebalance the influence of the rejected response, while FocalPO (Liu et al., 2025) adds a modulating factor to focus on correctly ranked ones. Anchored methods like APO (D'Oosterlinck et al., 2024) retain the DPO form but add explicit constraints to control the absolute likelihood drift of chosen or rejected responses. *HyPO* differs in where it acts: it preserves DPO's gradient form (same logistic weighting and β) but modifies the argument of the loss to $\Delta_\theta - \tilde{\Delta}\mathrm{ref}$ with $\tilde{\Delta}\mathrm{ref} = \max\{0, \Delta_{\mathrm{ref}}\}$, thereby reallocating gradient mass on pessimistic pairs without introducing an extra multiplicative weight.

3 Mitigating Mismatch in Direct Preference Alignment

To fully appreciate the design trade-offs in modern alignment methods, we briefly recall the KL-regularized RLHF principle (Christiano et al., 2017; Ziegler et al., 2019) that underpins them. We begin by considering a distribution of prompts $\mathcal{D}_p$ from which prompts x are sampled, and let π_θ be the trainable policy initialized from a supervised policy π_{SFT}. In KL-regularized policy optimization (e.g., PPO; Schulman et al., 2017), a canonical KL-regularized objective used in practice is:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}_p,\, y\sim\pi_\theta(\cdot|x)}\big[r_\phi(x,y)\big] \;-\; \tau\,\mathbb{E}_{x\sim\mathcal{D}_p}\big[D_{\text{KL}}\big(\pi_\theta(\cdot\mid x)\,\|\,\pi_{\text{ref}}(\cdot\mid x)\big)\big], \tag{1}$$

where r_ϕ is a reward model trained from human preferences, π_{ref} is a fixed reference (typically π_{SFT}), and $\tau > 0$ controls the regularization strength. For each x, $\pi_\theta^\star$ denotes the optimal policy, the per-context optimum of equation 1 has the Gibbs form (Peters et al., 2010; Azar et al., 2024):

$$\pi_\theta^\star(y\mid x) = \frac{\pi_{\text{ref}}(y\mid x)\,\exp\big(r_\phi(x,y)/\tau\big)}{\sum_{y'}\pi_{\text{ref}}(y'\mid x)\,\exp\big(r_\phi(x,y')/\tau\big)}, \tag{2}$$

which exposes the role of the KL term: it anchors π_θ near π_{ref}, making updates stable and discouraging distributional drift that could exploit reward misspecification (Schulman et al., 2015; 2017).

Direct Preference Optimization (DPO; Rafailov et al., 2023) circumvents explicit reward modeling and reinforcement learning by reformulating the objective for an offline setting, directly optimizing pairwise preferences (x, y^+, y^-) where $y^+ \succ y^-$ under a Bradley–Terry likelihood (Bradley & Terry, 1952). We define the policy and reference margins as:

$$\begin{aligned}\Delta_\theta(x,y^+,y^-) &= \log\pi_\theta(y^+\mid x) - \log\pi_\theta(y^-\mid x),\\ \Delta_{\text{ref}}(x,y^+,y^-) &= \log\pi_{\text{ref}}(y^+\mid x) - \log\pi_{\text{ref}}(y^-\mid x).\end{aligned} \tag{3}$$

DPO minimizes the logistic loss:

$$\mathcal{L}_{\text{DPO}} = \mathbb{E}_{(x,y^+,y^-)}\Big[-\log\sigma\Big(\beta\big(\Delta_\theta-\Delta_{\text{ref}}\big)\Big)\Big] = \mathbb{E}_{(x,y^+,y^-)}\Big[\log\Big(1+\exp\big(-\beta(\Delta_\theta-\Delta_{\text{ref}})\big)\Big)\Big], \tag{4}$$

where $\beta = 1/\tau$. Under the Bradley–Terry model, the first-order optimality condition shares the Gibbs structure of KL-regularized policy optimization (Rafailov et al., 2023) and satisfies:

$$\log\pi_\theta(y\mid x) - \log\pi_{\text{ref}}(y\mid x) = \beta\, r^\star(x,y) - \log Z(x). \tag{5}$$

This follows by taking logarithms of the KL-regularized optimum in Eq. 2:

$$\pi^\star(y\mid x) \;\propto\; \pi_{\text{ref}}(y\mid x)\,\exp\big(r^\star(x,y)/\beta\big), \tag{6}$$

where $r^\star$ is a preference-consistent reward and $Z(x)$ the partition function. This makes explicit how DPO preserves proximity to π_{ref} without explicit RL. Differentiating Eq. 4 with respect to the policy margin Δ_θ yields the per-example gradient-weight form:

$$\nabla_\theta\ell_{\text{DPO}} = -\beta\,\underbrace{\sigma\big(-\beta(\Delta_\theta-\Delta_{\text{ref}})\big)}_{w_{\text{DPO}}\in(0,1)}\,\nabla_\theta\Delta_\theta. \tag{7}$$

Although the introduction of the reference policy anchors the policy π_θ near π_{ref} via the relative margin $(\Delta_\theta - \Delta_{\text{ref}})$ serving as the gradient weight (Pan et al., 2025) in Eq. 7, thereby providing stability and suppressing distributional drift, it simultaneously introduces a *training–inference metric mismatch* (Chen et al., 2024; Meng et al., 2024). We refer to this side effect as *premature satisfaction*.

Premature Satisfaction. Let the scaled relative margin be:

$$z = \beta\big(\Delta_\theta - \Delta_{\text{ref}}\big), \qquad w_{\text{DPO}} = \sigma(-z) = \frac{1}{1+e^z}, \quad z \ge 0 \;\Rightarrow\; w_{\text{DPO}} \le e^{-z}. \tag{8}$$

Hence, from Eq. 7 we obtain: $\big\|\nabla_\theta\ell_{\text{DPO}}\big\| = \beta\,\sigma(-z)\,\big\|\nabla_\theta\Delta_\theta\big\| \;\le\; \beta\,e^{-z}\,\big\|\nabla_\theta\Delta_\theta\big\|$.

When the reference is pessimistic on a pair ($\Delta_{\text{ref}} < 0$), even if the current policy is still wrong in absolute terms ($\Delta_\theta < 0$), as soon as the relative margin becomes positive ($\Delta_\theta - \Delta_{\text{ref}} > 0$), the gradient shrinks exponentially, even though the policy remains wrong in absolute terms ($\Delta_\theta < 0$).

For example, take $\beta = 1$, $\Delta_{\text{ref}} = -3$, and $\Delta_\theta = -1$. Then $\Delta_\theta - \Delta_{\text{ref}} = 2$, we have $w_{\text{DPO}} = \sigma(-2) \approx 0.119$, $\big\|\nabla_\theta\ell_{\text{DPO}}\big\| \approx 0.119\,\big\|\nabla_\theta\Delta_\theta\big\|$. As illustrated by the lower-left region of the heatmap in Figure 1(a) (i.e., $\Delta_\theta - \Delta_{\text{ref}} > 0$), the weight w_{DPO} decays rapidly; when $\Delta_{\text{ref}} < 0$, this low-weight band still covers many points with $\Delta_\theta < 0$. *Premature satisfaction* reveals the tension between stability (proximality) and mismatch: while DPO suppresses drift, it may also systematically attenuate the training signal of examples that disagree with π_{ref} yet remain substantially wrong.

Reference-free Alignment. A natural reaction is to remove the reference to obtain an unbiased signal based on the absolute correctness of Δ_θ, aligned with inference-time likelihoods. Reference-free objectives (e.g., SimPO; Meng et al., 2024) optimize the absolute margin:

$$\mathcal{L}_{\text{abs}} = \mathbb{E}_{(x,y^+,y^-)}\Big[\,\ell_{\text{abs}}(\Delta_\theta)\,\Big], \qquad \ell_{\text{abs}}(\Delta_\theta) = \log\big(1 + \exp(-\beta\,\Delta_\theta)\big), \tag{9}$$

with gradient:

$$\frac{\partial \ell_{\text{abs}}}{\partial \Delta_\theta} = -\beta\,\sigma(-\beta\Delta_\theta), \qquad \nabla_\theta \ell_{\text{abs}} = -\beta\,\sigma(-\beta\Delta_\theta)\,\nabla_\theta\Delta_\theta. \tag{10}$$

Compared to Eq. 7, attenuation now depends only on the policy's own performance: pairs with $\Delta_\theta < 0$ inevitably receive large gradients until Δ_θ crosses zero, directly aligning the training signal with the inference-time objective (increasing $\log\pi_\theta(y^+ \mid x)$ and decreasing $\log\pi_\theta(y^- \mid x)$). However, Eq. 9 discards the very mechanism that implemented proximality in Eq. 4 and Eq. 5. Although recent work suggests that RLHF (Chu et al., 2025; Mukherjee et al., 2025) does not typically induce substantial parameter drift, removing this proximality nonetheless markedly reduces the robustness (Pan et al., 2025; Liu et al., 2024b) of reference-free alignment.

Direct Preference Optimization with Better Reference. Another approach argues that we should improve reference model to enhance DPO framework. This idea motivates using a stronger reference policy within the DPO framework (Liu et al., 2024b; Azar et al., 2024; Pan et al., 2025). The premise is that a more accurate reference margin, Δ_{ref}, makes the relative margin $(\Delta_\theta - \Delta_{\text{ref}})$ a better learning signal, which should improve optimization.

While often effective, these methods only modify the reference policy π_{ref} and leave the core DPO loss unchanged. Consequently, the phenomenon we term *premature satisfaction* persists whenever the reference is pessimistic on a pair (i.e., $\Delta_{\text{ref}} < 0$). To quantify this, we compute the sequence-level likelihood margin in Eq. equation 3 under three Llama-3-8B–based references (Base, SFT, and a SimPO-aligned model) on the UltraFeedback training split (Cui et al., 2023), and report (i) the density of Δ_{ref} and (ii) summary statistics (full settings and results in Appendix A). Stronger references are indeed pessimistic less often and shift the distribution of Δ_{ref} to the right; yet a substantial fraction, approximately 45%, of pairs remains in the pessimistic region even for the SimPO-aligned reference, which is designed to mitigate such mismatch. This imposes a practical ceiling on the "better reference" strategy: it improves stability and overall performance, but the *training–inference mismatch* persists wherever $\Delta_{\text{ref}} < 0$.

	Base	SFT	SimPO
Mean	-26.80	-22.65	14.15
p50	-7.02	-4.25	11.83

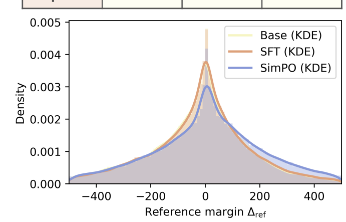


Figure 2: Distribution of the reference margin (Δ_{ref}) across different reference models. The table reports the mean and median (p50) of Δ_{ref} for each model.

3.1 Hybrid-DPO

In this section, we introduce *Hybrid-DPO* (*HyPO*), an approach designed to mitigate the *premature satisfaction* problem while preserving the stability of original DPO (Rafailov et al., 2023). The stability benefit of DPO, inherited from KL-regularized RLHF (Eq. 6), relies on the assumption that the reference policy π_{ref} provides a constructive learning signal. We argue that this KL-induced proximality should be *conditional*: When the reference is optimistic or neutral ($\Delta_{\text{ref}} \geq 0$), it provides a valid anchor for stable, relative improvement, as the DPO objective effectively preserves proximity to the reference (Eq. 5). However, when the reference is pessimistic ($\Delta_{\text{ref}} < 0$), it acts as a misleading prior that causes *premature satisfaction*; as analyzed in our discussion of the DPO gradient, the weight w_{DPO} in Eq. 7 attenuates even when the absolute margin Δ_θ remains negative. *HyPO* realizes this principle by clipping the reference margin: it preserves DPO's relative objective (relative margin, $(\Delta_\theta - \Delta_{\text{ref}})$) for reliable pairs but nullifies the reference's influence for pessimistic ones, compelling the model to optimize the absolute objective (absolute margin, Δ_θ) instead.

Objective. Formally, for a given threshold $\gamma \geq 0$ (with $\gamma = 0$ by default), we modify the DPO objective by replacing the reference margin Δ_{ref} with its clipped, non-negative counterpart:

$$\widetilde{\Delta}_{\text{ref}} = \max\{\Delta_{\text{ref}}, \gamma\}, \quad \text{and} \quad \mathcal{L}_{\text{HyPO}} = \mathbb{E}\left[\log\left(1 + \exp\left(-\beta(\Delta_\theta - \widetilde{\Delta}_{\text{ref}})\right)\right)\right]. \tag{11}$$

The gradient of the per-example loss retains the DPO structure, weighted by the modified margin:

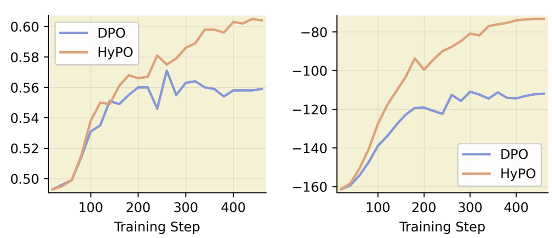
$$\nabla_\theta \mathcal{L}_{\text{HyPO}} = -\beta \cdot \underbrace{\sigma\left(-\beta(\Delta_\theta - \widetilde{\Delta}_{\text{ref}})\right)}_{w_{\text{HyPO}}} \cdot \nabla_\theta \Delta_\theta. \tag{12}$$

By construction, *HyPO* coincides with DPO on *non-pessimistic* samples ($\Delta_{\text{ref}} \geq \gamma$), i.e., its per-example weight in Eq. 12 equals that of DPO in Eq. 7 ($w_{\text{HyPO}} = w_{\text{DPO}}$), thereby preserving proximal stability where the reference is reliable. On pessimistic samples ($\Delta_{\text{ref}} < \gamma$), clipping nullifies the misleading anchor and restores an absolute–margin–driven update, pointwise dominating the reference-free weight $w_{\text{abs}} := \sigma(-\beta\Delta_\theta)$ induced by Eq. 10 ($w_{\text{HyPO}} \geq w_{\text{abs}}$), thus preventing premature attenuation. An intuitive visualization is shown in Figure 1(c).

For a globally smooth objective, we can replace the hard maximum with a softplus function (Dugas et al., 2000). The smoothed reference margin is defined as:

$$\widetilde{\Delta}_{\text{ref}} = \gamma + \frac{1}{\alpha}\log\left(1 + \exp(\alpha(\Delta_{\text{ref}} - \gamma))\right), \quad \alpha > 0, \tag{13}$$

where α controls the smoothness ($\alpha \to \infty$ recovers Eq. 11). This simple, smoothed formulation allows *HyPO* to be implemented as a plug-in modification to the standard DPO loss.



(a) Agreement Rate (b) Absolute Margin

Pairwise Win Rate (%): Columns vs. Rows

Win Rate (%)	SFT	DPO	HyPO (Ours)
SFT	-	15.7	**12.9**
DPO	84.3	-	**44.1**
HyPO (Ours)	**87.1**	**55.9**	-

(c) Win-Rate Matrix

Figure 3: *HyPO* improves inference-aligned evaluation metrics and pairwise win rates. (a) *Absolute agreement rate* over training (higher is better). (b) *Absolute margin on the pessimistic subset* ($\Delta_{\text{ref}} < 0$). (c) *Pairwise win-rate*. Each cell is the win rate (%) of the row model against the column model on AlpacaEval 2.0 (Li et al., 2023). All results use the same SFT checkpoint of Llama-3-8B-Base (AI@Meta, 2024) trained on UltraFeedback (Cui et al., 2023) with either DPO or our *HyPO*; the training/evaluation pipeline and optimization hyperparameters are identical and set to the DPO configuration from Zephyr (Tunstall et al., 2023). See Section 4.1 for more settings.

Effectiveness of *HyPO*. To quantify how effectively the *HyPO* objective mitigates the *training–inference mismatch* compared to DPO (Rafailov et al., 2023), we track two key metrics on the evaluation set that are closely aligned with inference-time performance. The first, the *Absolute Agreement Rate*, serves as a global measure of overall performance. Defined as $\text{Agree}@t := \Pr[\Delta_\theta^{(t)} > 0]$ (with Δ_θ the absolute policy margin as in Eq. 3), it measures the fraction of evaluation pairs whose absolute likelihood ordering matches the desired preference, directly mirroring the inference-time decision rule. Further, we use a more targeted *Absolute Margin on the Pessimistic Subset*, which tracks the average absolute margin, $\mathbb{E}[\Delta_\theta \mid \Delta_{\text{ref}} < 0]$, on the subset where the (fixed) reference policy is pessimistic. This metric isolates the challenging cases that trigger *premature satisfaction* in DPO and precisely measures *HyPO*'s ability to improve learning on them. Further details on metric definitions are provided in Appendix E. To directly compare final model outputs, we report a pairwise win rate matrix using AlpacaEval 2.0 (Li et al., 2023). For models i (row) and j (column), the entry $W_{i,j} := 100 \times \Pr[\text{model } i \succ \text{model } j]$ is the percentage of prompts on which the row model's response is preferred to the column model's, under the same judge and prompts; diagonals are undefined. Configuration and evaluation protocol are detailed in Section 4.1.

As demonstrated in Figure 3(a) and (b), *HyPO* achieves a significantly faster and higher increase in both the global agreement rate and the pessimistic subset margin, confirming its effectiveness in mitigating the *training–inference mismatch*. In this run, the final agreement rate increases from 55.9% for DPO to 60.4% for *HyPO*. As shown in Figure 3(c), *HyPO* outperforms *DPO* head-to-head ($W_{\text{HyPO,DPO}} = 55.9\%$ vs. $W_{\text{DPO,HyPO}} = 44.1\%$). Moreover, against the common baseline *SFT*, both alignment methods win, with *HyPO* achieving the larger advantage.

A key advantage of *HyPO* is that it keeps a DPO-like form and is therefore compatible with enhancements for the DPO. In our practical implementation, we leverage this compatibility in two ways. First, we adopt a better reference (Liu et al., 2024b; Pan et al., 2025) by using a pre-aligned, same-family model instead of the SFT default. Second, we apply a uniform home advantage margin h to impose stronger learning pressure, leading to the final objective: $\Delta_\theta \; - \; \max\{\Delta_{\text{ref}},\, \gamma\} \; - \; h$. We do not use length normalization (Park et al., 2024) due to its hyperparameter sensitivity.

4 Experiments

We conduct a series of experiments to validate the efficacy of *HyPO*. We first establish its outperformance against a set of offline preference alignment baselines on standard benchmarks. Subsequently, we perform targeted analyses and ablations to probe the robustness and scalability of *HyPO*.

4.1 Experimental Setup and Results

To ensure comprehensive and fair comparison of our proposed *HyPO* against various direct preference alignment baselines, we closely follow the established experimental settings from existing works (Meng et al., 2024; Rashidinejad & Tian, 2025; Gorbatovski et al., 2025). More detailed experimental settings can be found in Appendix A.

Models and datasets settings. Our main experiments evaluate two open-source model families, Mistral-7B (Jiang et al., 2023) and Llama3-8B (AI@Meta, 2024), using the UltraFeedback dataset ($\approx$ 61k training samples; Cui et al., 2023 for preference alignment. We conduct these experiments in two distinct settings. In the *Base setting*, we first perform supervised fine-tuning (SFT) on the raw pretrained models (e.g., Mistral-7B-v0.1, Meta-Llama-3-8B) using the UltraChat-200k dataset ($\approx$ 200k training samples; Ding et al., 2023). The resulting SFT policy then serves as the initial policy for preference alignment on UltraFeedback. In the *Instruct setting*, we bypass the SFT step and directly apply our preference alignment methods to the official instruction-tuned models (e.g., Mistral-7B-Instruct-v0.2, Meta-Llama-3-8B-Instruct), again using the UltraFeedback dataset.

Training settings. Our training methodology is grounded in established practices from recent alignment literature to ensure robustness and fair comparison (Tunstall et al., 2023; 2025). For the initial SFT stage, we use a peak learning rate of 2×10^{-5}. For each reproduced direct preference alignment method, we conduct a small grid search over hyperparameters: the peak learning rate is selected from $\{5 \times 10^{-7}, 1 \times 10^{-6}\}$, while the DPO temperature β is chosen from $\{0.01, 0.1\}$ for standard objectives or from $\{2, 10\}$ for methods employing length normalization. For our proposed *HyPO*, we use a checkpoint of the same architecture pre-aligned with SimPO. The home advantage margin was set to $h = 10$, while the threshold was kept at its default of $\gamma = 0$. Detailed settings are listed in Appendix A. All models are trained for a single epoch using the AdamW optimizer (Loshchilov & Hutter, 2017), and we employ a cosine learning rate schedule with a warm-up phase over the first 10% of training steps and a global batch size of 128.

Evaluation benchmarks. We primarily assess model performance on two popular instruction-following benchmarks: AlpacaEval 2.0 (Li et al., 2023; Dubois et al., 2024) and Arena-Hard-v0.1 (Li et al., 2024a;b). Consistent with prior work (Rashidinejad & Tian, 2025; Gorbatovski et al., 2025), we use GPT-4-Preview-1106 as the automated judge to compute the primary metric: the win rate, which is the percentage of times a model's response is preferred over a baseline's. The specific baseline model varies by benchmark: AlpacaEval reports win rates against GPT-4-Preview-1106 itself, whereas Arena-Hard uses GPT-4-0314. Our evaluation protocols follow established practices from Tunstall et al. (2023); Meng et al. (2024); Tunstall et al. (2025), with the minor modification of using a more recent vLLM release (Kwon et al., 2023) for generation.

Direct preference alignment baselines. We compare our method against a suite of direct preference alignment baselines. Our comparison starts with two classic off-policy methods: SLiC-HF (Zhao et al., 2023) and DPO (Rafailov et al., 2023), the latter of which serves as our primary baseline.

CPO (Xu et al., 2024) and KTO (Ethayarajh et al., 2024) are DPO variants that reshape the loss from, respectively, a contrastive perspective and an asymmetric utility. SimPO (Meng et al., 2024) is a high-performing reference-free representative that operates on the absolute likelihood margin. FocalPO (Liu et al., 2025), TR-DPO (Gorbatovski et al., 2025), and RainbowPO (Zhao et al., 2025) are recent DPO improvements: focal reweighting, trust-region constraints on the relative margin for stability, and a unified practical recipe combining length normalization and policy mixing.

Table 1: Main results reported from AlpacaEval 2.0 (Li et al., 2023; Dubois et al., 2024) and Arena-Hard-v0.1 (Li et al., 2024a;b). LC and WR denote *length-controlled win rate* and *raw win rate*, respectively. The best results are highlighted in bold, and the second-best are underlined.

Method	**Mistral-Base (7B)**			**Mistral-Instruct (7B)**		
	AlpacaEval		**Arena-Hard**	**AlpacaEval**		**Arena-Hard**
	LC (%)	**WR (%)**	**WR (%)**	**LC (%)**	**WR (%)**	**WR (%)**
SLiC-HF (Zhao et al., 2023)	11.6	9.1	5.4	32.4	31.2	16.9
DPO (Rafailov et al., 2023)	22.6	18.5	7.9	35.1	31.4	15.4
CPO (Xu et al., 2024)	13.1	11.6	6.4	34.9	39.9	<u>21.0</u>
KTO (Ethayarajh et al., 2024)	12.9	9.3	6.6	35.0	31.3	17.5
SimPO (Meng et al., 2024)	27.3	25.4	<u>11.2</u>	<u>38.4</u>	<u>40.0</u>	20.5
FocalPO (Liu et al., 2025)	25.8	19.7	8.2	35.9	35.0	18.7
TR-DPO (Gorbatovski et al., 2025)	24.9	21.4	9.5	36.5	33.7	18.2
RainbowPO (Zhao et al., 2025)	<u>28.4</u>	<u>26.7</u>	9.2	35.7	33.9	18.2
HyPO (Ours)	**32.8**	**29.6**	**13.9**	**38.9**	**47.9**	**25.2**
Method	**Llama-3-Base (8B)**			**Llama-3-Instruct (8B)**		
	AlpacaEval		**Arena-Hard**	**AlpacaEval**		**Arena-Hard**
	LC (%)	**WR (%)**	**WR (%)**	**LC (%)**	**WR (%)**	**WR (%)**
SLiC-HF (Zhao et al., 2023)	19.8	15.9	14.3	36.7	36.8	25.1
DPO (Rafailov et al., 2023)	24.3	21.9	23.0	40.9	41.3	31.5
CPO (Xu et al., 2024)	22.3	24.6	12.2	38.1	40.4	30.0
KTO (Ethayarajh et al., 2024)	23.6	20.3	18.4	40.5	39.0	30.5
SimPO (Meng et al., 2024)	30.7	26.2	30.1	46.0	43.1	32.1
FocalPO (Liu et al., 2025)	27.2	25.4	27.9	45.1	<u>43.6</u>	30.2
TR-DPO (Gorbatovski et al., 2025)	<u>31.8</u>	<u>30.2</u>	<u>31.0</u>	<u>46.7</u>	42.7	<u>32.5</u>
RainbowPO (Zhao et al., 2025)	30.3	27.1	28.6	<u>46.7</u>	43.5	31.3
HyPO (Ours)	**34.7**	**33.6**	**33.5**	**49.5**	**46.2**	**35.2**

Main Results on Benchmarks. We present the main results of our comparative analysis in Table 1. Across all four experimental settings, including two model families (Mistral-7B and Llama-3-8B) and two initial states (Base and Instruct), our proposed method, *HyPO*, consistently and significantly outperforms all baselines on both AlpacaEval and Arena-Hard benchmarks. The results show that *HyPO* achieves an average relative improvement of 41.2% over the original DPO. Furthermore, when compared to SimPO, a strong reference-free competitor, *HyPO* delivers an average relative improvement of 15.1%. Taken together, these results highlight the practical value of our methodology. By mitigating the reference mismatch while preserving the core DPO framework, our method translates a targeted theoretical improvement into superior performance on benchmarks.

4.2 Further Analysis and Ablations

To further validate the effectiveness and robustness of *HyPO*, we evaluate downstream transfer, run ablations (including sensitivity to γ), measure training runtime, and study scaling. For comparability, we benchmark against one reference-based baseline (DPO; Rafailov et al., 2023) and one reference-free baseline (SimPO; Meng et al., 2024) under largely the same settings as Section 4.1.

Downstream tasks. One of the challenges in alignment is that aligning a model can reduce a model's general capabilities. This degradation is often attributed to the over-optimization (Gao et al., 2023; Rafailov et al., 2024) of the alignment objective. To assess whether our approach preserves broad

Table 2: Ablation study of *HyPO*'s components.

Llama-3-Models	Base		Instruct		
	LC(%)	WR(%)	LC(%)	WR(%)	Avg. Δ(%)
HyPO (Ours)	34.7	33.6	49.5	46.2	-
- *Home Advantage*	33.2	32.5	47.4	45.1	- 1.5
- *Better Reference*	29.3	27.1	45.3	45.9	- 4.1
- *BR and HA*	28.8	28.5	45.2	44.9	- 4.2
Standard DPO	24.3	21.9	40.9	41.3	- 8.9

Figure 4: Sensitivity to the threshold γ, using Meta-Llama-3-8B-Instruct.

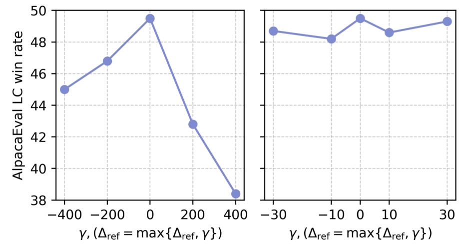


Table 3: Further results reported from AlpacaEval 2.0 (Li et al., 2023; Dubois et al., 2024) and Arena-Hard-v0.1 (Li et al., 2024a;b). LC and WR denote *length-controlled win rate* and *raw win rate*, respectively. The best results are highlighted in bold.

Method	**Helpsteer2**			**Mistral-Nemo-Instruct (12B)**			**Qwen-2.5-7B**		
	AlpacaEval		**Arena-Hard**	**AlpacaEval**		**Arena-Hard**	**AlpacaEval**		**Arena-Hard**
	LC (%)	**WR (%)**	**WR (%)**	**LC (%)**	**WR (%)**	**WR (%)**	**LC (%)**	**WR (%)**	**WR (%)**
DPO (Rafailov et al., 2023)	16.2	14.3	4.5	50.4	49.2	35.5	27.9	22.4	28.8
SimPO (Meng et al., 2024)	18.6	16.0	7.0	52.1	46.4	33.9	33.7	22.2	33.8
HyPO (Ours)	**22.3**	**19.5**	**9.3**	**55.7**	**54.9**	**38.9**	**38.0**	**30.7**	**36.2**

utility, we evaluated the aligned models on a comprehensive suite of downstream tasks, using LM Evaluation Harness (Biderman et al., 2024). Our evaluation covers four key domains. We test knowledge and language understanding using MMLU (Hendrycks et al., 2020) and ARC-Challenge (Clark et al., 2018). For commonsense reasoning, we employ HellaSwag (Zellers et al., 2019) and Winogrande (Sakaguchi et al., 2021). Factuality is assessed with TruthfulQA (Lin et al., 2021), and mathematical reasoning is measured by GSM8K (Cobbe et al., 2021). The results, presented in Table 5 (see Appendix D for full details), demonstrate that *HyPO* maintains or improves performance across all evaluated tasks relative to the corresponding baselines. This indicates that our method avoids additional performance degradation on downstream tasks that can be induced by alignment.

Ablations and threshold γ sensitivity. We ablate two enhancements in our *HyPO* implementation: the use of a better reference model and a home advantage margin. The results in Table 2 show that both components are beneficial. Notably, the ablated *HyPO* variants still significantly outperform the standard DPO baseline, a finding consistent with our head-to-head comparisons in Figure 3, confirming the effectiveness of our core mechanism. Furthermore, we analyze the model's sensitivity to the threshold γ from Eq. 11. As shown in Figure 4, performance remains stable across a range of γ values around 0. Given this low sensitivity, we adopt the most principled and interpretable setting, $\gamma = 0$. These validate the contributions of our proposed method and the stability of our configuration.

Running time. As a drop-in replacement for standard DPO, *HyPO* introduces negligible additional computation. In the experiment shown in Figure 3, conducted with identical hardware and settings, the wall clock training time for *HyPO* differed from that of DPO by $\approx 1\%$ (7.1 hours vs. 7.2 hours).

Example Responses. To offer a qualitative view, Appendix C compares sample model outputs. The responses suggest that *HyPO* can generate more considered answers that acknowledge subtleties in the prompt, distinguishing it from the more direct outputs of baseline methods.

Scalability. To further validate the scalability of our proposed method, we conduct extended experiments across varying scales, datasets, and architectures. We evaluate *HyPO* along three axes: (i) a dataset shift from UltraFeedback to HelpSteer2 (Wang et al., 2024b); (ii) a model-size shift to Mistral-Nemo-12B (Mistral AI, 2024); and (iii) an architectural shift to Qwen-2.5-7B (Qwen, 2025). We follow the evaluation protocol and most hyperparameters from Section 4.1 (details in Appendix A). As summarized in Table 3, *HyPO* consistently improves over baselines across all settings, indicating robustness to distribution changes, parameter scaling, and diverse model families.

5 CONCLUSION

In this paper, we introduced *Hybrid-DPO (HyPO)* to address *premature satisfaction*, a critical failure in Direct Preference Optimization (DPO) where a pessimistic reference model down-weights learning signals. *HyPO*'s simple conditional mechanism, clipping the reference margin at zero, corrects this bias while retaining DPO's inherent stability. Empirically, *HyPO* consistently and significantly outperforms both various DPO and strong reference-free baselines across a range of models and benchmarks, proving robust and scalable without an additional "alignment tax". Our work provides evidence that direct preference alignment could be enhanced by conditionally debiasing the reference signal, rather than discarding it, thereby achieving a more effective and stable compromise.

6 LIMITATIONS AND FUTURE WORK

Our approach aims to *mitigate* rather than fully resolve the training-inference mismatch. By design, *HyPO* retains the reference-based behavior on optimistic pairs to preserve training stability and leverage better reference models, meaning the training–inference mismatch persists for this subset. Furthermore, our method assumes general label reliability; in cases of severe label noise where the reference model correctly disagrees with a wrong label, *HyPO* enforces a strong learning signal and may arguably amplify noise compared to DPO. Finally, while we prioritized a minimalist, parameter-free clipping mechanism, more sophisticated functional forms, such as adaptive clipping thresholds or learned gating mechanisms, remain promising directions for future research.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their insightful and constructive comments. Suqin Yuan thanks Muyang Li and Li He for their valuable advice. Tongliang Liu is partially supported by the following Australian Research Council projects: FT220100318, DP260102466, DP220102121, LP220100527, LP220200949. Suqin Yuan is supported in part by the OpenAI Researcher Access Program (projects 05950 and 19000). This research/project is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-003). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

ETHICS STATEMENT

Our research aims to improve the alignment of large language models with human preferences, a goal intended to enhance their safety, helpfulness, and reliability. We primarily use publicly available datasets, such as UltraFeedback (Cui et al., 2023) and HelpSteer2 (Wang et al., 2024b), which are aggregated from existing sources. While we did not collect new human data, we acknowledge that these datasets may contain inherent biases or sensitive information reflective of their origins. Our work does not aim to create new models from scratch but to refine existing ones (Mistral and Llama-3; Jiang et al., 2023; AI@Meta, 2024). We recognize that any advancement in AI capabilities carries a dual-use risk (Hong et al., 2024b; Huang et al., 2024; Lin et al., 2025; Zhang et al.; Huang et al., 2026). However, the explicit goal of our method, *HyPO*, is to make models' behavior more predictable and aligned with positive human values, thereby mitigating risks associated with misaligned AI. We believe that research into robust alignment techniques is a crucial step toward the responsible development and deployment of AI systems.

REPRODUCIBILITY STATEMENT

Our method, Hybrid-DPO (HyPO), is designed as a minimalist, plug-in modification to the standard Direct Preference Optimization (DPO) objective. This design ensures that our work can be easily reproduced by integrating our logic into popular alignment libraries and codebases that already support DPO, such as Alignment-handbook (Tunstall et al., 2025). All models and datasets used in our experiments are open-source and publicly accessible through the Hugging Face, ensuring full transparency. The base models include Mistral-7B (Jiang et al., 2023), Meta-Llama-3-8B (AI@Meta, 2024), and Mistral-Nemo-12B (Mistral AI, 2024). The datasets for supervised fine-tuning and preference alignment include UltraChat-200k (Ding et al., 2023), UltraFeedback (Cui et al., 2023), HelpSteer2 (Wang et al., 2024b), and OpenAssistant2 (Köpf et al., 2023). For evaluation on benchmarks like AlpacaEval 2.0 (Li et al., 2023; Dubois et al., 2024), Arena-Hard-v0.1 (Li et al., 2024a;b), and LM Evaluation Harness (Biderman et al., 2024), we followed established protocols that utilize the GPT-4-Preview-1106 API as an automated judge. While access to this commercial API is managed by a third party, our evaluation methodology is public and can be adapted for use with other capable judge models. We have release our code to facilitate further research: `https://github.com/tmllab/2026_ICLR_HyPO`.

The Use of Large Language Models

During the preparation of this manuscript, the authors utilized large language models (LLMs) in several capacities. As writing and coding assistants, they were used to improve grammar, spelling, and sentence structure for clarity and to generate code for training and visualization. The subject of this research is the alignment of large language models themselves, by using LLMs such as Mistral and Llama-3 (Jiang et al., 2023; AI@Meta, 2024). Furthermore, as part of the evaluation methodology, the authors employed a closed-source LLM's API (GPT-4-Preview-1106) to act as an automated judge for assessing the quality of the trained LLMs. The authors maintained full intellectual control of this paper. LLMs were used as tools to augment the research process, and the final manuscript reflects the authors' own work and insights.

References

AI@Meta. Llama 3 model card. 2024.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

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

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 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.

Angelica Chen, Sadhika Malladi, Lily H Zhang, Xinyi Chen, Qiuyi Zhang, Rajesh Ranganath, and Kyunghyun Cho. Preference learning algorithms do not learn preference rankings. *Advances in Neural Information Processing Systems*, 37:101928–101968, 2024.

Jay Hyeon Cho, JunHyeok Oh, Myunsoo Kim, and Byung-Jun Lee. Rethinking dpo: The role of rejected responses in preference misalignment. *arXiv preprint arXiv:2506.12725*, 2025.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust dpo: Aligning language models with noisy feedback. *arXiv preprint arXiv:2403.00409*, 2024.

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

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. 2023.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*, 2023.

Karel D'Oosterlinck, Winnie Xu, Chris Develder, Thomas Demeester, Amanpreet Singh, Christopher Potts, Douwe Kiela, and Shikib Mehri. Anchored preference optimization and contrastive revisions: Addressing underspecification in alignment. *arXiv preprint arXiv:2408.06266*, 2024.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, 13, 2000.

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

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Alexey Gorbatovski, Boris Shaposhnikov, Alexey Malakhov, Nikita Surnachev, Yaroslav Aksenov, Ian Maksimov, Nikita Balagansky, and Daniil Gavrilov. Learn your reference model for real good alignment. *arXiv preprint arXiv:2404.09656*, 2025.

Li He, Qiang Qu, He Zhao, Stephen Wan, Dadong Wang, Lina Yao, and Tongliang Liu. Unifying stable optimization and reference regularization in RLHF. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QpqBqCTtW4.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024a.

Ziming Hong, Zhenyi Wang, Li Shen, Yu Yao, Zhuo Huang, Shiming Chen, Chuanwu Yang, Mingming Gong, and Tongliang Liu. Improving non-transferable representation learning by harnessing content and style. In *The twelfth international conference on learning representations*, 2024b.

Zhuo Huang, Chang Liu, Yinpeng Dong, Hang Su, Shibao Zheng, and Tongliang Liu. Machine vision therapy: Multimodal large language models can enhance visual robustness via denoising in-context learning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=LwOfVWgEzS.

Zhuo Huang, Qizhou Wang, Ziming Hong, Shanshan Ye, Bo Han, and Tongliang Liu. Is gradient ascent really necessary? memorize to forget for machine unlearning. *arXiv preprint arXiv:2602.06441*, 2026.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023.

Andreas Köpf, Yannic Kilcher, Dimitri Von Rütte, Sotiris Anagnostidis, Zhi Rui Tam, Keith Stevens, Abdullah Barhoum, Duc Nguyen, Oliver Stanley, Richárd Nagyfi, et al. Openassistant conversations-democratizing large language model alignment. *Advances in neural information processing systems*, 36:47669–47681, 2023.

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

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024a.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. From live data to high-quality benchmarks: The arena-hard pipeline, April 2024b.

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, 5 2023.

Runqi Lin, Alasdair Paren, Suqin Yuan, Muyang Li, Philip Torr, Adel Bibi, and Tongliang Liu. Force: Transferable visual jailbreaking attacks via feature over-reliance correction. *arXiv preprint arXiv:2509.21029*, 2025.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

Aiwei Liu, Haoping Bai, Zhiyun Lu, Yanchao Sun, Xiang Kong, Simon 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. *arXiv preprint arXiv:2410.04350*, 2024a.

Tong Liu, Xiao Yu, Wenxuan Zhou, Jindong Gu, and Volker Tresp. FocalPO: Enhancing preference optimizing by focusing on correct preference rankings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, July 2025.

Yixin Liu, Pengfei Liu, and Arman Cohan. Understanding reference policies in direct preference optimization. *arXiv preprint arXiv:2407.13709*, 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

Mistral AI. Mistral-nemo-12b-instruct, 2024. Model card; Apache-2.0; accessed 2025-09-21.

Sagnik Mukherjee, Lifan Yuan, Dilek Hakkani-Tur, and Hao Peng. Reinforcement learning finetunes small subnetworks in large language models. *arXiv preprint arXiv:2505.11711*, 2025.

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

Junshu Pan, Wei Shen, Shulin Huang, Qiji Zhou, and Yue Zhang. Pre-dpo: Improving data utilization in direct preference optimization using a guiding reference model. *arXiv preprint arXiv:2504.15843*, 2025.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.

Jan Peters, Katharina Mulling, and Yasemin Altun. Relative entropy policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 24, pp. 1607–1612, 2010.

Qwen. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115v2*, 2025.

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. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sushil Sikchi, Joey Hejna, Brad Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *Advances in Neural Information Processing Systems*, 37:126207–126242, 2024.

Paria Rashidinejad and Yuandong Tian. Sail into the headwind: Alignment via robust rewards and dynamic labels against reward hacking. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3505–3506, 2020.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897, Lille, France, 07–09 Jul 2015. PMLR.

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, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Archit Sharma, Sedrick Scott Keh, Eric Mitchell, Chelsea Finn, Kushal Arora, and Thomas Kollar. A critical evaluation of ai feedback for aligning large language models. *Advances in Neural Information Processing Systems*, 37:29166–29190, 2024.

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. Mapo: Advancing multilingual reasoning through multilingual alignment-as-preference optimization. *arXiv preprint arXiv:2401.06838*, 2024.

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

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Shengyi Huang, Kashif Rasul, Alvaro Bartolome, Carlos M. Patiño, Alexander M. Rush, and Thomas Wolf. The alignment handbook, 2025.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*, 2024a.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. Helpsteer 2: Open-source dataset for training top-performing reward models. *Advances in Neural Information Processing Systems*, 37:1474–1501, 2024b.

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

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. $beta$-dpo: Direct preference optimization with dynamic $beta$. *Advances in Neural Information Processing Systems*, 37:129944–129966, 2024b.

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 *International Conference on Machine Learning*, 2025.

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. *arXiv preprint arXiv:2401.08417*, 2024.

Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Robust llm alignment via distributionally robust direct preference optimization. *arXiv preprint arXiv:2502.01930*, 2025.

Xiliang Yang, Feng Jiang, Qianen Zhang, Lei Zhao, and Xiao Li. Dpo-shift: Shifting the distribution of direct preference optimization. *arXiv preprint arXiv:2502.07599*, 2025.

Suqin Yuan, Lei Feng, and Tongliang Liu. Late stopping: Avoiding confidently learning from mislabeled examples. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16079–16088, 2023a.

Suqin Yuan, Lei Feng, and Tongliang Liu. Early stopping against label noise without validation data. In *The Twelfth International Conference on Learning Representations*, 2024.

Suqin Yuan, Lei Feng, Bo Han, and Tongliang Liu. Enhancing sample selection against label noise by cutting mislabeled easy examples. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Suqin Yuan, Runqi Lin, Lei Feng, Bo Han, and Tongliang Liu. Instance-dependent early stopping. In *The Thirteenth International Conference on Learning Representations*, 2025b.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*, 2024.

Yichi Zhang, Siyuan Zhang, Yao Huang, Zeyu Xia, Zhengwei Fang, Xiao Yang, Ranjie Duan, Dong Yan, Yinpeng Dong, and Jun Zhu. Stair: Improving safety alignment with introspective reasoning. In *Forty-second International Conference on Machine Learning*.

Hanyang Zhao, Genta Indra Winata, Anirban Das, Shi-Xiong Zhang, David Yao, Wenpin Tang, and Sambit Sahu. RainbowPO: A unified framework for combining improvements in preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 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. *arXiv preprint arXiv:2406.11827*, 2024.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A Experimental Details

This section provides a comprehensive overview of the experimental settings used to validate our proposed method, *Hybrid-DPO (HyPO)*. Our experimental design is grounded in established practices from recent alignment literature to ensure fair, robust, and reproducible comparisons (Tunstall et al., 2023; Meng et al., 2024; Tunstall et al., 2025). We detail the models and datasets, the multi-stage training pipeline, hyperparameter tuning, and the evaluation protocols for all experiments.

A.1 Models and Datasets

Our experiments leverage a suite of publicly available models and datasets to ensure full transparency and facilitate replication. We conduct experiments on two popular open-source model families: Mistral-7B (Jiang et al., 2023) and Meta-Llama-3-8B (AI@Meta, 2024). We test each model in two settings: a *Base* setting, starting from the raw pretrained weights, and an *Instruct* setting, starting from the official instruction-tuned checkpoints. For our scalability study, we also use Mistral-Nemo-Instruct-12B (Mistral AI, 2024).

Supervised Fine-Tuning (SFT) Datasets. For experiments in the *Base* setting, we first conduct supervised fine-tuning to obtain a capable initial policy. We use UltraChat-200k (Ding et al., 2023) for the main experiments and OpenAssistant2 (Köpf et al., 2023) for the dataset transfer experiment.

Preference Datasets. For the core preference alignment stage, our main experiments use the UltraFeedback dataset (Cui et al., 2023), from which we use the official training split for optimization and hold out a fixed validation set. For the dataset transfer scalability experiment, we use the HelpSteer2 dataset (Wang et al., 2024b). In *Instruct* setting, we use UltraFeedback dataset released by Meng et al. (2024).

A.2 Pipeline and Hyperparameters

Our training process consists of two main stages for the *Base* setting (SFT followed by preference alignment) and one stage for the *Instruct* setting (preference alignment only).

Supervised Fine-Tuning (SFT) Stage.

For the *Base* models, we first create an SFT version by training on the relevant dataset for one epoch. We use the AdamW optimizer (Loshchilov & Hutter, 2017) with a global batch size of 128, a peak learning rate of 2×10^{-5}, and a cosine learning rate schedule with a 10% warm-up phase. The maximum sequence length is set to 2048. This SFT model serves as both the initial policy and the default reference model (π_{ref}) for preference alignment.

Preference Alignment Stage.

We conduct a small targeted hyperparameter search for each method.

General Hyperparameters. To ensure consistent optimization dynamics, across all preference alignment experiments, we train for a single epoch with a global batch size of 128, a maximum sequence length of 2048, and a maximum prompt length of 1800 tokens. We use the AdamW optimizer (Loshchilov & Hutter, 2017) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a cosine learning rate schedule with a 10% warm-up phase. To maintain the fixed global batch size of 128 despite varying memory requirements across stages (e.g., the additional reference model overhead in DPO) and hardware configurations (ranging from 1 to 4 GPUs), we adjusted the per-device micro-batch size and gradient accumulation steps based on the available GPU memory and model scale. For any specific set of comparative experiments, we ensured that these effective batch size settings were strictly identical across all baselines to guarantee a fair comparison.

Method-Specific Hyperparameters. We identified the peak learning rate and the DPO temperature β as the most sensitive hyperparameters. For each method we reproduced, we performed a small grid search and selected the best configuration based on the lowest validation loss on our held-out set. In our experiments, we used some of the checkpoints released by Meng et al. (2024).

- *Peak Learning Rate:* Searched over $\{5 \times 10^{-7}, 1 \times 10^{-6}\}$.

- *Temperature* β: For standard objectives (DPO, CPO, KTO, FocalPO, TR-DPO, and HyPO.), we searched over $\{0.01, 0.1\}$. For methods employing length normalization (e.g., RainbowPO, SimPO), we followed Meng et al. (2024) recommendations and searched over $\{2, 10\}$.

Evaluation Stage.

Instruction-Following Benchmarks. We use GPT-4-Preview-1106 as the automated judge. The primary metric is the win rate (%). For AlpacaEval 2.0 (Li et al., 2023; Dubois et al., 2024), we report both the raw win rate (WR) and the length-controlled win rate (LC WR) against the GPT-4-Preview-1106 baseline model. Following Meng et al. (2024); Tunstall et al. (2023), we employed a sampling-based decoding approach. The temperature was set to 0.7 for the Mistral-Base (Jiang et al., 2023) setting, 0.5 for the Mistral-Instruct models (7B; Jiang et al., 2023/12B; Mistral AI, 2024), and 0.9 across both Llama-3-8B (AI@Meta, 2024). For Arena-Hard-v0.1 (Li et al., 2024a;b), we report the win rate against the GPT-4-0314 baseline model, and we use greedy decoding.

Downstream Task Evaluation. To assess whether alignment methods introduce an *alignment tax* by degrading general capabilities, we use the LM Evaluation Harness (Biderman et al., 2024). We evaluate performance on MMLU (5-shot) (Hendrycks et al., 2020), ARC-Challenge (25-shot) (Clark et al., 2018), HellaSwag (10-shot) (Zellers et al., 2019), Winogrande (5-shot) (Sakaguchi et al., 2021), TruthfulQA (0-shot) (Lin et al., 2021), and GSM8K (5-shot) (Cobbe et al., 2021).

A.3 Implementation and infrastructure details

Our implementation builds on the open-source Alignment-handbook codebase (Tunstall et al., 2025), which itself is built around Hugging Face Transformers and the `trl` library. HyPO is realized as a thin modification of `trl`'s `DPOTrainer` (see Appendix G); all remaining components (data loading, logging, checkpointing, evaluation scripts) are reused from the Alignment-handbook stack with minimal changes. See Appendix A.3 for low-level implementation details.

We train all models with DeepSpeed (Rasley et al., 2020) ZeRO and bfloat16 mixed precision. For the main runs, we use ZeRO stage-1 with CPU offloading of optimizer states, which substantially reduces GPU memory usage while keeping the implementation simple. On a subset of configurations, we also ran ZeRO stage-3 without CPU offloading and observed that all reported metrics (loss, win-rate) match well. Based on this observation, we adopted the stage-1 + CPU-offload configuration for all large-scale sweeps reported in the paper. All main results are obtained on compute nodes equipped with four NVIDIA H100 96GB GPUs. We have verified that the same configuration can be run on two H100 GPUs, or on a single H100 with CPU offloading, at the cost of longer wall-clock time but essentially identical performance Gradient checkpointing is enabled for all models to further reduce memory consumption. Unless otherwise stated, we use Flash Attention 2 (Dao, 2023) for faster attention computation, bfloat16 mixed-precision for accelerated throughput, and realized via data parallelism and gradient accumulation.

A.4 Analysis of Reference Model Pessimism (Figure 2)

To motivate our work, the analysis in Figure 2 quantifies the prevalence of pessimistic reference margins ($\Delta_{\mathrm{ref}} < 0$). This experiment was conducted on the full training split of the UltraFeedback dataset (Cui et al., 2023). We computed the sequence-level log-likelihood margins for three distinct reference models derived from the Meta-Llama-3-8B family: (i) the raw base model, (ii) our SFT checkpoint trained on UltraChat-200k (use Zephyr settings; Tunstall et al., 2023), and (iii) a model further aligned from the SFT checkpoint using SimPO (use SimPO settings; Meng et al., 2024). For each pair in the dataset, we calculated Δ_{ref} without length normalization. We provide summary statistics for each distribution, including the mean and median (p50). The results confirmed that even for a strong, pre-aligned reference model, a substantial fraction of pairs ($\approx 45\%$) remained in the pessimistic region, underscoring the persistence of the training-inference mismatch.

A.5 Direct DPO vs. HyPO Comparison (Figure 3)

Figure 3 presents a controlled comparison between standard DPO and our proposed *HyPO*. The goal of this experiment was to isolate the impact of our conditional reference mechanism. Both methods started from the exact same SFT checkpoint of Llama-3-8B-Base and were trained on UltraFeedback

(Cui et al., 2023). Critically, all hyperparameters were held identical and were based on the DPO configuration from Zephyr (Tunstall et al., 2023), including a learning rate of 5×10^{-7} and $\beta = 0.01$. For this specific comparison, ***HyPO*** was used in its most basic form: it used the SFT model as its reference (not a better reference) and was configured with $\gamma = 0$ and no home advantage margin ($h = 0$). This setup ensures a fair, one-to-one comparison against the DPO baseline. The metrics in Figure 3(a) and 3(b) were computed periodically on our held-out validation set throughout the single training epoch. The win-rate matrix in Figure 3(c) was generated by evaluating the final SFT, DPO, and HyPO checkpoints against each other on AlpacaEval (Li et al., 2023).

B Analysis of DPO with SFT Loss

In this section, we explore a natural baseline: adding a supervised fine-tuning (SFT) loss on the chosen response to the standard DPO objective. This acts as a regularizer to ensure the model maintains high likelihood for preferred responses. The DPO + SFT-on-chosen objective is defined as:

$$\mathcal{L}_{\text{DPO + SFT-on-chosen}} = \mathcal{L}_{\text{DPO}} + \lambda \, \mathcal{L}_{\text{SFT}}(y_w), \tag{14}$$

where $\mathcal{L}_{\text{SFT}}(y_w) = -\log \pi_\theta(y_w \mid x)$.

B.1 Hyperparameter Search

To ensure a fair comparison, we conducted a hyperparameter sweep for the coefficient λ over the range $\{0.01, 0.03, 0.1, 0.3, 0.5\}$. We established $\lambda = 0.5$ as the upper bound for our sweep because we observed that, at this value, the magnitude of the SFT loss term was approximately equal to the DPO loss term at the beginning of training. We posit that the auxiliary SFT loss should act as a regularizer rather than the dominant training signal; therefore, λ should not exceed the point where SFT dominates the preference optimization. Empirically, we found that the optimal performance on AlpacaEval 2.0 was achieved at $\lambda = 0.03$. We use this best-performing configuration for the definitive comparison below.

B.2 Comparison and Analysis

We compared the optimized DPO + SFT-on-chosen ($\lambda = 0.03$) against the standard DPO and our HyPO method. In addition to win rates, we measured the Agreement Rate (defined in Appendix E) to assess how well the model's likelihood ordering aligns with the ground truth preferences.

Table 4: Comparison of DPO, DPO + SFT-on-chosen ($\lambda = 0.03$), and HyPO. All models are Llama-3-Base-8B initialized from the same SFT checkpoint.

Method	**AlpacaEval LC (%)**	**AlpacaEval WR (%)**	**Agreement Rate (%)**
DPO (Baseline)	24.3	21.9	55.9
DPO + SFT-on-chosen	26.1	21.9	56.3
HyPO (Ours)	**28.8**	**28.5**	**60.4**

As shown in Table 4, DPO + SFT-on-chosen ($\lambda = 0.03$) yields only a marginal improvement in win rate (+1.8% LC) and Agreement Rate (+0.4%). HyPO significantly outperforms this baseline (+4.5% LC over baseline).

C Examples Responses

The following is an example question along with responses from various models from the Alpaca Eval 2.0 helpful subset base pipeline. HyPO response shows analytical thinking, a deeper reasoning process, and consideration of implicit factors.

1. gpt4-1106-preview: Most comprehensive and beginner-safe; actionable progressions.
2. HyPO (ours): Well-structured with reps/warmup/cooldown; a few choices (Russian twists, Pallof press) may be advanced or equipment-dependent.
3. SimPO: Solid variety and set guidance; minor glitch (stray char) and several advanced core picks for true beginners.
4. DPO: Long single list with multiple advanced moves (hanging leg raises); fewer regressions and weaker structure.

Responses from Different Models

Question

What are the best exercises for beginners?

GPT-4-Preview-1106 (Baseline): For beginners, the best exercises are those that are easy to learn, safe, and effective at building strength, endurance, and flexibility. Here is a list of exercises that are generally well-suited for beginners, covering different aspects of fitness:

1. **Cardiovascular Exercises:**
   - Walking: Start with a brisk walk and gradually increase your pace and distance.
   - Cycling: Begin with a flat route before tackling hills.
   - Swimming: Try a few laps with rest in between, gradually increasing your endurance.
   - Jumping Jacks: A simple full-body exercise to get your heart rate up.
2. **Strength Training Exercises:**
   - Bodyweight Squats: Focus on form and depth before adding weight.
   - Push-ups: Start with wall or knee push-ups if a full push-up is too difficult.
   - Planks: Begin by holding the position for short periods and gradually increase the time.
   - Lunges: Practice stationary lunges to master the form before attempting walking lunges.
   - Dumbbell Exercises: Use light weights to perform exercises like bicep curls, shoulder presses, and tricep extensions.
3. **Flexibility and Balance Exercises:**
   - Stretching: Incorporate basic stretches for all major muscle groups, holding each stretch for 15-30 seconds.
   - Yoga: Basic poses like Child's Pose, Downward-Facing Dog, and Warrior can improve flexibility and balance.
   - Tai Chi: This gentle martial art is great for balance and stress reduction.
4. **Core Strengthening Exercises:**
   - Bird-Dog: This exercise improves core stability and coordination.
   - Dead Bug: Another great move for core strength that's easy on the back.
   - Seated Russian Twists: Begin without weight and focus on the twisting motion.
5. **Functional Exercises:**
   - Step-ups: Use a low step at first and focus on form.
   - Farmer's Walk: Carry a light weight in each hand and walk for a set distance, maintaining good posture.
   - Glute Bridges: These help strengthen the glutes and stabilize the core.

It's essential for beginners to focus on proper form to prevent injuries and to start with a level of intensity that matches their current fitness level. It's also advisable to consult with a fitness professional or a physician before starting any new exercise program, especially if you have pre-existing health conditions or concerns.

Responses from Different Models

Question

What are the best exercises for beginners?

DPO: As a beginner, it's essential to start with exercises that are safe, effective, and enjoyable to help you build a strong foundation and momentum in your fitness journey. Here are some of the best exercises for beginners:

1. **Squats:** Works your legs, glutes, and core. Stand with feet shoulder-width apart, then bend knees and lower your body down, keeping back straight.
2. **Push-ups:** Targets chest, shoulders, and triceps. Start in a plank position with hands shoulder-width apart, lower your body until your chest almost touches the ground, then push back up.
3. **Lunges:** Works legs, glutes, and lower body. Stand with feet together, take a large step forward with one foot, and lower your body down, keeping back knee almost touching the ground. Alternate legs.
4. **Planks:** Strengthens core and improves posture. Hold a plank position for 30-60 seconds, engaging your abs and keeping your body straight from head to heels.
5. **Chest Press (using dumbbells or resistance bands):** Works chest, shoulders, and triceps. Lie on a flat surface or sit with knees bent, holding weights or resistance bands, and press them upwards towards the ceiling.
6. **Rows (using dumbbells or resistance bands):** Targets back and biceps. Hold weights or resistance bands and pull them towards your chest, keeping your elbows close to your body.
7. **Bicycle Crunches:** Engages core and improves obliques. Lie on your back with hands behind your head, alternate bringing each elbow towards the opposite knee, as if pedaling a bicycle.
8. **Leg Raises (hanging leg raises or leg lifts):** Works abs and lower back. Hang from a pull-up bar or lie on your back with arms extended, lifting your legs straight up towards the ceiling.
9. **Wall Sit:** Strengthens legs and glutes. Stand with back against a wall, slide down into a seated position with knees bent at a 90-degree angle, and hold for 30 seconds to 1 minute.
10. **Brisk Walking or Jogging:** Improves cardiovascular fitness and burns calories. Start with short sessions and gradually increase duration and intensity.

Remember to:

- Start slow and focus on proper form and technique.
- Begin with lighter weights and gradually increase the load as you build strength.
- Warm up before exercising with 5-10 minutes of light cardio and stretching.
- Listen to your body and rest when needed.
- Consult with a healthcare professional or fitness expert if you have any concerns or injuries.

As you progress, you can incorporate more exercises and intensity to challenge yourself further. Happy exercising!

Responses from Different Models

Question

What are the best exercises for beginners?

SimPO: As a beginner, it's essential to start with exercises that are safe, effective, and manageable to build a strong foundation for your fitness journey. Here are some of the best exercises for beginners:

1. **Bodyweight exercises:**
   - Push-ups: Works chest, shoulders, and triceps.
   - Squats: Targets legs, glutes, and core.
   - Lunges: Works legs, glutes, and improves balance.
   - Planks: Engages core muscles for stability and endurance.
   - Pull-ups (or assisted pull-ups): Targets back, shoulders, and arms.
2. **Resistance band exercises:**
   - Banded squats, lunges, and chest presses: Add variety to bodyweight exercises with resistance bands.
3. **Cardio exercises:**
   - Brisk walking: Easy to start and improves cardiovascular fitness.
   - Jogging or running (begin with short intervals): Increases endurance and burns calories.
   - Swimming or cycling: Low-impact alternatives for those with joint concerns.
4. **Upper body exercises with light weights or dumbbells:**
   - Dumbbell chest press: Targets chest and shoulders.
   - Bicep curls: Works biceps.
   - Tricep dips (using a chair or bench): Targets triceps.
5. **Core exercises:**
   - Russian twists: Engages obliques.
   - Leg raises: Targets lower abs.
   - Bicycle crunches: Works entire core.

Remember to:

1. Start slowly and gradually increase intensity and duration as you become more comfortable.
2. Focus on proper form and technique to avoid injuries.
3. Begin with 2-3 sets of 8-12 repetitions for each exercise.
4. Listen to your body and rest when needed.
5. Consult with a healthcare professional or fitness expert to create a personalized workout plan.

These exercises will help you build a solid foundation in strength, flexibility, and cardiovascular endurance. As you progress, you can always add more variety and intensity to your routine.

Responses from Different Models

Question

What are the best exercises for beginners?

HyPO (ours): As a beginner, it's essential to start with exercises that are gentle, easy to learn, and progressive in intensity to build a strong foundation and prevent injury. Here are some of the best exercises for beginners:

1. **Bodyweight exercises:**
   - Push-ups
   - Squats
   - Lunges
   - Planks
   - Bicycle crunches (modified or regular)
   - Wall sit
2. **Resistance band exercises:**
   - Banded squats and lunges
   - Banded chest press (using a resistance band or light dumbbells)
   - Banded rows (using a resistance band)
3. **Low-impact cardio:**
   - Brisk walking
   - Jogging in place
   - Swimming (if accessible)
   - Cycling (stationary or outdoor)
4. **Upper body strength exercises with light weights:**
   - Dumbbell rows
   - Shoulder presses (using light weights or resistance bands)
   - Bicep curls
   - Tricep dips (using a chair or bench)
5. **Core exercises:**
   - Leg raises (lying or hanging)
   - Russian twists (using a medicine ball or weight plate)
   - Pallof press (using a cable machine or resistance band)

Remember to:

1. Start slow and gentle, with 2-3 sets of 8-12 repetitions for each exercise.
2. Focus on proper form and technique over the number of repetitions or weight used.
3. Listen to your body and rest when needed.
4. Gradually increase intensity and volume as you become more comfortable with the exercises.
5. Incorporate exercises that work multiple muscle groups at once (e.g., squats engage legs, glutes, and core).

Additionally, consider the following general tips:

- Consult with a healthcare professional or fitness expert to create a personalized workout plan.
- Warm up before exercising with 5-10 minutes of light cardio and dynamic stretching.
- Cool down afterwards with static stretches to help prevent soreness and improve flexibility.

By starting with these beginner-friendly exercises, you'll set yourself up for a strong foundation in fitness and enjoy a positive exercise experience.

D Downstream Tasks

One of the primary challenges in preference alignment is the potential for a model's general capabilities to degrade, a phenomenon often attributed to over-optimization on the alignment objective (Gao et al., 2023; Rafailov et al., 2024). To rigorously assess whether our approach introduces any additional "alignment tax" and preserves broad utility, we evaluated aligned models on a comprehensive suite of downstream tasks. For comparability, we benchmark against one reference-based baseline (DPO; Rafailov et al., 2023) and one reference-free baseline (SimPO; Meng et al., 2024). All three methods are evaluated starting from the same SFT checkpoint, with identical prompt formatting and decoding hyperparameters.

The evaluation was conducted using the standard LM Evaluation Harness framework (Biderman et al., 2024), with inference performed via the Hugging Face Transformers library. We follow the default task implementations in LM Evaluation Harness and report accuracy (or exact-match for GSM8K) as provided by the framework. We measured performance across four key domains, using the following few-shot settings for each task:

- *Knowledge & Language Understanding:* MMLU (Hendrycks et al., 2020) (5-shot) and ARC-Challenge (Clark et al., 2018) (25-shot).
- *Commonsense Reasoning:* HellaSwag (Zellers et al., 2019) (10-shot) and Winogrande (Sakaguchi et al., 2021) (5-shot).
- *Factuality:* TruthfulQA (Lin et al., 2021) (0-shot).
- *Mathematical Reasoning:* GSM8K (Cobbe et al., 2021) (5-shot).

The results, presented in Table 5, demonstrate that *HyPO* consistently performs on par with or better than the initial model and other alignment methods. Notably, *HyPO* achieves the highest average score in three out of the four experimental settings. This indicates that our method successfully aligns with human preferences without incurring a significant penalty on the model's core capabilities.

Table 5: Downstream task evaluation results, using LM Evaluation Harness (Biderman et al., 2024). The best results are highlighted in bold, and the second-best are underlined.

	MMLU	ARC	HellaSwag	TruthfulQA	Winogrande	GSM8K	Average
Mistral-Base (7B)							
Initial Model	59.0	54.2	60.8	28.0	77.7	34.6	52.4
DPO (Rafailov et al., 2023)	57.6	58.7	64.4	40.3	77.4	31.5	<u>55.0</u>
SimPO (Meng et al., 2024)	58.4	58.6	63.6	33.1	78.1	35.0	54.5
HyPO (Ours)	58.5	60.2	63.5	39.1	78.9	31.8	**55.3**
Mistral-Instruct (7B)							
Initial Model	59.2	58.7	66.1	52.6	78.1	44.1	59.8
DPO (Rafailov et al., 2023)	58.8	60.9	67.6	51.2	78.3	41.6	59.7
SimPO (Meng et al., 2024)	58.1	62.5	67.5	55.0	77.4	39.8	<u>60.1</u>
HyPO (Ours)	59.6	66.8	67.4	54.0	77.6	40.6	**61.0**
Llama-3-Base (8B)							
Initial Model	63.7	56.9	61.4	31.5	77.4	50.3	56.9
DPO (Rafailov et al., 2023)	63.3	61.6	64.7	37.9	78.0	54.7	<u>60.0</u>
SimPO (Meng et al., 2024)	62.1	63.4	64.5	38.0	77.2	50.4	59.3
HyPO (Ours)	63.6	65.4	64.3	42.8	78.9	49.7	**60.8**
Llama-3-Instruct (8B)							
Initial Model	65.8	56.4	59.0	36.1	77.3	74.8	61.6
DPO (Rafailov et al., 2023)	65.6	61.3	60.6	42.4	76.4	73.5	**63.3**
SimPO (Meng et al., 2024)	64.7	59.8	56.0	43.8	72.5	71.3	61.4
HyPO (Ours)	65.0	64.4	58.7	45.3	73.3	70.3	<u>62.8</u>

E Evaluation Metric Definitions

To provide a fine-grained analysis of model performance during training, as shown in Figure 3, we track two key inference-aligned metrics on a held-out evaluation set of preference pairs (x, y_w, y_l), where $y_w \succ y_l$. These metrics are designed to directly quantify how effectively our proposed method, *HyPO*, mitigates the training-inference mismatch inherent in DPO. Both metrics are computed using the policy and reference log-likelihood margins, Δ_θ and Δ_{ref}, as formally defined in the main text in Equation 3. We denote the policy margin at a given training checkpoint t as $\Delta_\theta^{(t)}$.

Absolute Agreement Rate. The Absolute Agreement Rate provides a global measure of how well the policy's log-likelihoods align with the ground-truth preferences. It is defined as the probability that the policy assigns a higher log-likelihood to the winning response y_w than the losing response y_l:

$$\text{Agree}\,@t := \Pr_{(x,y_w,y_l)}\left[\Delta_\theta^{(t)}(x, y_w, y_l) > 0\right].$$

At inference time, a model's performance depends on its *absolute* log-likelihoods, as the reference model is not used. This metric therefore directly reflects the desired inference-time behavior. A higher agreement rate signifies that the policy is more accurately ranking responses in absolute terms, indicating a successful reduction of the training-inference mismatch.

Absolute Margin on the Pessimistic Subset. This metric is a targeted diagnostic designed to measure performance on the specific subset of data where DPO is most prone to failure. We first define the *pessimistic subset* as all evaluation pairs where the reference model incorrectly prefers the losing response:

$$\mathcal{D}_{\text{pessimistic}} := \{(x, y_w, y_l) \mid \Delta_{\text{ref}}(x, y_w, y_l) < 0\}.$$

The metric is then the average policy margin, Δ_θ, computed exclusively over this subset:

$$\text{Margin}_{\text{pessimistic}}@t := \mathbb{E}_{(x,y_w,y_l)\in\mathcal{D}_{\text{pessimistic}}}\left[\Delta_\theta^{(t)}(x, y_w, y_l)\right].$$

This metric directly probes the phenomenon of *premature satisfaction*. For pairs in $\mathcal{D}_{\text{pessimistic}}$, DPO's learning signal weakens as soon as the policy margin Δ_θ surpasses the negative reference margin Δ_{ref}, even if Δ_θ itself is still negative. A strong and increasing positive value for this metric demonstrates that the policy is successfully applying corrective pressure and overcoming the misleading signal from the pessimistic reference, a core goal of *HyPO*.

F Analysis of the Smoothness Parameter α

In Eq. 13 we introduced a smooth variant of the clipped reference margin using a softplus transformation, with smoothness parameter α:

$$\widetilde{\Delta}_{\text{ref}} = \gamma + \frac{1}{\alpha}\log\bigl(1 + \exp(\alpha(\Delta_{\text{ref}} - \gamma))\bigr), \quad \alpha > 0,$$

where smaller values of α yield a smoother transition, while $\alpha \to \infty$ recovers the hard $\max\{\Delta_{\text{ref}}, \gamma\}$.

To assess the sensitivity of *HyPO* to this smoothness parameter, we ran a small sweep over $\alpha \in \{1, 10, 100\}$ using Meta-Llama-3-8B-Instruct aligned on UltraFeedback, evaluated on AlpacaEval 2.0. All other settings are identical to those in our main experiments.

Table 6: Sensitivity of *HyPO* to the smoothness parameter α in Eq. 13 on Meta-Llama-3-8B-Instruct.

Method	α	AlpacaEval LC (%)	AlpacaEval WR (%)
HyPO	1	48.3	45.5
HyPO (default)	10	49.5	46.2
HyPO	100	49.6	46.2

The results in Table 6 show that *HyPO* is robust to the choice of α in a reasonable range: $\alpha = 10$ and $\alpha = 100$ yield nearly identical performance, while a much smoother transition ($\alpha = 1$) causes only a modest degradation. This supports our default choice of $\alpha = 10$ as a simple and effective setting, and suggests that the benefits of *HyPO* do not rely on fine-tuning this parameter.

G Minimal code change from DPO to HyPO

We implement HyPO by modifying only the computation of the DPO logits inside the original `DPOTrainer.dpo_loss` in `trl`. Let

$$\Delta_\theta = \log \pi_\theta(y_c \mid x) - \log \pi_\theta(y_r \mid x), \qquad \Delta_{\text{ref}} = \log \pi_{\text{ref}}(y_c \mid x) - \log \pi_{\text{ref}}(y_r \mid x),$$

be the policy and reference log-ratios. Standard DPO uses $\Delta_\theta - \Delta_{\text{ref}}$ as the logit input to the sigmoid loss. HyPO simply replaces Δ_{ref} with a clipped (or smoothed) reference margin $\widetilde{\Delta}_{\text{ref}}$:

$$\widetilde{\Delta}_{\text{ref}} = \max(\Delta_{\text{ref}}, \gamma) \quad \text{or} \quad \widetilde{\Delta}_{\text{ref}} = \gamma + \tau \, \text{softplus}\left(\frac{\Delta_{\text{ref}} - \gamma}{\tau}\right),$$

where $\tau = 1/\alpha$, and uses $\Delta_\theta - \widetilde{\Delta}_{\text{ref}}$ in the DPO loss.

The concrete code change in `DPOTrainer.dpo_loss` is:

```
# Original DPO (simplified):
pi_logratios = policy_chosen_logps - policy_rejected_logps
if self.reference_free:
    ref_logratios = torch.zeros_like(pi_logratios)
else:
    ref_logratios = reference_chosen_logps - reference_rejected_logps

logits = pi_logratios - ref_logratios
```

is replaced by the following HyPO version:

```
# HyPO: Conditional Reference Clipping
pi_logratios = policy_chosen_logps - policy_rejected_logps
if self.reference_free:
    ref_logratios = torch.zeros_like(pi_logratios)
else:
    ref_logratios = reference_chosen_logps - reference_rejected_logps

# ---------------- HyPO Modification Start -----------------
gamma = torch.tensor(self.args.hypo_gamma, device=ref_logratios.device)

if getattr(self.args, "hypo_tau", 0.0) > 0.0:
    # Smooth HyPO: Eq. 13 (using tau for temperature)
    tau = self.args.hypo_tau
    ref_logratios = gamma + tau * F.softplus((ref_logratios - gamma) / tau
        )
else:
    # Hard HyPO: Eq. 11 (standard max)
    ref_logratios = torch.maximum(ref_logratios, gamma)
# ----------------- HyPO Modification End ------------------

logits = pi_logratios - ref_logratios
```

In our experiments, we introduce *hypo_gamma* (scalar threshold γ, default 0) and optionally *hypo_tau* (smoothing factor $\tau = 1/\alpha$) as hyperparameters.