From Implicit Rewards to Explicit Objectives for Human Preference Alignment
STOR 743 - Final Report

Minji Kim, Akshay Sakanaveeti, Nikos Dimou*

Department of Statistics and Operations Research
University of North Carolina at Chapel Hill

1 Introduction

LLMs power diverse interactive applications, yet they are fundamentally trained to predict the next token in a sequence. However, they often fail to capture user intent. Raw pretraining maximizes statistical probability, whereas deployment requires helpfulness, safety, and alignment with human preferences. This the main motivation for Hu et al. [10], on which this report is based on. In this report, we discuss how to *fine-tune* an LLM so that its behavior matches human preferences, given only pairwise comparisons between candidate responses as feedback. Concretely, for a prompt x, humans are shown two responses y_w and y_l and asked "which one do you prefer?"; from many such comparisons, we wish to construct a new policy $\pi_\theta(y \mid x)$ whose responses are more likely to be preferred.

This problem arises in the rapidly expanding literature on reinforcement learning from human feedback (RLHF) [1, 12, 16]. RLHF views human preferences as noisy observations of a reward function and optimizes an RL-style objective in which the model is an agent, generation is a sequence of actions, and human feedback defines a return that must be balanced against staying close to a reference policy. This motivation is formalized by the regularized objective

$$\max_{\pi_\theta} \mathbb{E}_{x,y\sim\pi_\theta(\cdot|x)}[r(x,y)] - \lambda\,\mathbb{E}_x\big[\mathrm{KL}\big(\pi_\theta(\cdot \mid x) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid x)\big)\big]\,,$$

which captures the central tradeoff between reward improvement and preservation of prior behavior.

Although inspired by RL, preference-based fine-tuning behaves more like an offline contextual bandit: the environment is a human annotator or reward model, numeric rewards are not directly observed, and data are collected from a fixed reference policy rather than through online exploration. The EXPO paper highlights that many widely used methods replace full RL with static surrogate losses optimized by SGD.

The classical RLHF pipeline trains a reward model from human comparisons (often with a Bradley–Terry likelihood) and then applies PPO to optimize the above objective. This separation is conceptually clear but computationally heavy and can suffer from reward hacking,

*mkim5@unc.edu, sakshay@unc.edu, dimou@unc.edu

distributional shift, and instability under misspecification [14, 17]. These challenges motivated Direct Preference Optimization (DPO) [13], which encodes reward implicitly through the ratio of the policy to the reference model and reduces training to a single offline loss.

EXPO [10] examines whether such implicit methods behave consistently with the RL formulation. The authors argue that DPO and related quasi-convex objectives fail to preserve regions where the reference model is already optimal and collapse to a deterministic best-response policy as $\lambda \to 0$, rather than approaching a full preference-optimal distribution. EXPO instead proposes explicit objectives that maintain the reward-plus-KL structure while grounding everything in observable preference distributions. Their analysis shows that EXPO preserves optimal behavior, interpolates smoothly between π_{ref} and a preference-optimal policy, and performs competitively with DPO in alignment benchmarks.

Many existing approaches (DPO, PPO, IPO, GPO, f-DPO) rely on a single global scalar loss that applies identical regularization everywhere. This can lead to structural failures such as loss of diversity or instability under hyperparameter changes. EXPO is motivated by addressing these issues directly rather than modifying the RLHF formulation.

RL provides conceptual clarity by defining optimality through a balance of human preference and trust in the reference model, and it explains the appearance of KL regularization. However, full RLHF with PPO is often unnecessary for static datasets and remains vulnerable to reward misspecification. Implicit methods such as DPO are efficient but can behave unpredictably from an RL perspective. EXPO aims to combine the scalability of these implicit approaches with guarantees that more faithfully reflect the structure of the underlying RL-inspired objective.

2 Implicit Reward Structures: Strengths and Limitations

Implicit reward methods simplify the alignment process by reparameterizing the RL objective into a standard supervised loss. As a result, methods such as DPO and its variants admit closed-form, fully offline objectives that can be optimized without explicit policy sampling or reward modeling. At a high level, these implicit reward methods share a common structure: preference learning is reduced to minimizing an expectation over pairwise comparison data,

$$\mathcal{L}_{\text{QPO}}(\pi_\theta) = \mathbb{E}_{(x,y_w,y_l,s)\sim\mathcal{D}}\left[\ell\Big(g\big(x, y_w, y_l;\, \pi_\theta, \pi_{\text{ref}}\big)\Big)\right], \tag{2.1}$$

where g is a scalar function of π_θ and π_{ref} (often a log-ratio), and ℓ is a monotone penalty function. The EXPO paper refers to this broad class as quasi-convex preference optimization (QPO): ℓ is assumed to be differentiable and quasi-convex, so it increases monotonically away from its minimum [8]. Under this viewpoint, DPO, IPO, GPO [18], and f-DPO [19] are all special cases of (2.1), differing only in the precise choice of g and ℓ. They share some practical advantages:

- They avoid training a separate reward model.
- They reduce preference learning to minimizing a single offline loss on $\mathcal{D}$.
- They make it easy to reuse standard supervised-learning infrastructure.

Despite these advantages, the EXPO paper argues that by mathematically tying the policy strictly to an implicit reward, these methods suffer from sub-optimal regularization. The core

issue is that QPO-based objectives apply uniform regularization pressure across all inputs, failing to distinguish between prompts where the reference model is already performing well and those where it needs improvement.

Failure to Preserve Optimal Policies. Consider a partition of the prompt space into a set d_x^{good}, where the reference model coincides with the optimal preference policy π^*, i.e., $\pi_{\text{ref}} = \pi^*$, and a set d_x^{bad}, where it does not. Intuitively, an effective alignment method should act selectively: it should shift the policy towards human preferences on "bad" prompts (d_x^{bad}) while preserving the reference model's behavior on "good" prompts (d_x^{good}) where it is already optimal. However, Theorem 2.1 demonstrates that QPO methods cannot achieve this balance. See Appendix A for the proof.

Theorem 2.1. *Given the prompt partitioning, reference policy, and optimal policy described above, define* $\hat{\pi}_\theta^{QPO} := \arg\min_{\pi_\theta} \ell_{QPO}(\pi_\theta, \pi_{ref}, \psi, \mu, \lambda)$ *for any fixed selection of* $\{\psi, \mu, \lambda\}$. *Then under relatively mild assumptions on the labeled responses in* $\mathcal{D}_{tr}$, *if* $dist(\hat{\pi}_\theta^{QPO}, \pi^*) < dist(\pi_{ref}, \pi^*)$ *for* $x \in d_x^{bad}$, *then* $dist(\hat{\pi}_\theta^{QPO}, \pi^*) > 0$ *for* $x \in d_x^{good}$.

3 Explicit Preference Optimization (EXPO)

To address the shortcomings mentioned above, the authors suggest an alternative approach *Explicit Preference Optimization (EXPO)* that targets human preferences directly without relying on implicit reward reparameterization.

3.1 Bradley-Terry Optimality

Standard preference optimization methods, such as RLHF and DPO, suffer from a critical failure mode in the unregularized limit ($\lambda \to 0$). Specifically, these methods converge to a deterministic, mode-collapsed policy π^δ that exclusively outputs the single highest-reward response. For example, given three responses where y_3 is preferred over y_2 with probability 0.6, a mode-collapsed policy will generate y_3 with probability 1.0 and y_2 with probability 0.0. This behavior creates two fundamental issues:

(1) **False Certainty:** By assigning zero mass to valid alternative responses, the model effectively assumes the preference gap is infinite, misrepresenting the noisy nature of human signals.

(2) **Loss of Diversity:** Valid, high-quality responses are eliminated entirely, reducing the generative capability of the model.

To address this, the authors propose that an ideal generative model should reflect the *gradations of human preference* rather than collapsing to a single point. Therefore, we define the target optimal policy π^* as the **Bradley-Terry (BT) optimal policy**. Unlike the greedy policy π^δ, the BT-optimal policy ensures that the relative frequency of generating two responses matches their pairwise preference strength. Formally, a policy π^* is BT-optimal if it satisfies the condition:

$$p^*(y_1 > y_2 \mid x) = \frac{\pi^*(y_1 \mid x)}{\pi^*(y_1 \mid x) + \pi^*(y_2 \mid x)} \tag{3.1}$$

3.2 EXPO model

To practically achieve the Bradley-Terry optimal policy π^* when $\lambda \to 0$, the authors propose a compositional loss function that decouples preference learning from regularization. Unlike RLHF or DPO, which implicitly maximize a reward proxy, this objective explicitly targets the preference distribution.

The total loss is defined as a weighted sum of a supervised preference term and an unsupervised regularization term [10]:

$$l^c_{EXPO}(\pi_\theta, \pi_{ref}, \lambda) = l_{sup}(\pi_\theta) + \lambda l_{unsup}(\pi_\theta, \pi_{ref}) \tag{3.2}$$

3.2.1 The Supervised Term (l_{sup})

The supervised term is designed to maximize the likelihood that the policy reproduces human pairwise preferences. It is derived by minimizing the KL divergence between the ground-truth preference distribution p^* and the policy's induced preference distribution p_θ:

$$l_{sup}(\pi_\theta) = \mathbb{E}_{\{y_w, y_l, x\} \sim \mathcal{D}_{tr}} \left[\log \left(1 + \frac{\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)} \right) \right] \tag{3.3}$$

Crucially, minimizing this term alone recovers the BT-optimal distribution π^*. This is because l_{sup} is derived by minimizing the KL divergence between the ground-truth preference distribution p^* and the distribution p_θ induced by the policy:

$$l_{sup}(\pi_\theta) = \mathbb{E}_{\{y_1, y_2\} \sim \pi_{ref}, x \sim \mathcal{D}_x} \left[\mathrm{KL}(p^*(z|y_1, y_2, x) || p_\theta(z|y_1, y_2, x)) \right] \tag{3.4}$$

where the induced preference probability p_θ is defined as:

$$p_\theta(z = 1|y_1, y_2, x) = \frac{\pi_\theta(y_1|x)}{\pi_\theta(y_1|x) + \pi_\theta(y_2|x)} \tag{3.5}$$

A key property of this formulation is that minimizing l_{sup} uniquely aligns the induced preference p_θ with the ground truth p^*. Consequently, the optimal solution to this unregularized term naturally recovers the BT-optimal distribution π^*.

3.2.2 The Unsupervised Term (l_{unsup})

To ensure the model retains the capabilities of the pre-trained model, the unsupervised term applies standard regularization anchored to the reference policy:

$$l_{unsup}(\pi_\theta, \pi_{ref}) = \mathbb{E}_{x \sim \mathcal{D}_x} \left[\mathrm{KL}(\pi_{ref}(y|x) || \pi_\theta(y|x)) \right] \tag{3.6}$$

Next we discuss that above optimisation procedure also provides a resolution to the limitation mentioned in Section 2

3.3 Preserving Optimal Policies

A critical limitation of QPO methods is the inability to improve policy performance in regions where the reference model is poor without degrading performance in regions where it is already optimal. Proposition 4.2 asserts that EXPO overcomes this limitation through its decoupled loss structure.

Proposition 3.1 (Preservation of Optimality)**.** Assume the prompt distribution decomposes into disjoint sets d_x^{good} and d_x^{bad}. Let the reference policy be optimal in the good region ($\pi_{ref} = \pi^*$ for $x \in d_x^{good}$) but suboptimal in the bad region ($dist(\pi_{ref}, \pi^*) \gg 0$). The EXPO minimizer $\hat{\pi}_\theta^{EXPO}$ satisfies:

(1) $\hat{\pi}_\theta^{EXPO}(y|x) = \pi^*(y|x)$ for all $x \in d_x^{good}$ (Perfect Preservation).

(2) $dist(\hat{\pi}_\theta^{EXPO}, \pi^*) < dist(\pi_{ref}, \pi^*)$ for $x \in d_x^{bad}$ (Selective Improvement).

Proof Sketch. The total EXPO loss can be decomposed linearly based on the prompt regions:

$$l_{EXPO} = \mathbb{E}[l(\pi_\theta, x)\mathbb{I}\{x \in d_x^{good}\}] + \mathbb{E}[l(\pi_\theta, x)\mathbb{I}\{x \in d_x^{bad}\}] \tag{3.7}$$

In the good region where $\pi_{ref} = \pi^*$, both the supervised preference term l_{sup} (which targets π^*) and the unsupervised term l_{unsup} (which targets π_{ref}) are simultaneously minimized at the same global optimum π^*. Consequently, the gradient in the good region is zero at π^*, allowing the optimizer to reduce the loss in d_x^{bad} without requiring a trade-off that compromises the optimal solution in d_x^{good} as in 2.1. □

4 Numerical Experiments: Interpretation and Validation

The empirical section of the EXPO paper is designed to probe how different preference-optimization objectives behave under controlled conditions. This aligns well with the aim of this report: the goal is to understand when an RL-style alignment framework is appropriate, what advantages it offers over simpler baselines, and which structural properties of the learning objective matter most in practice. Rather than reproducing the numerical experiments of [10], we instead offer some narrative and interpretation of the plots provided, and we compare the performance of EXPO with that of DPO, IPO and f-DPO. The experiments are organized around three questions:

(1) In simple synthetic environments where the "true" preference distribution is known, do the different losses actually converge to the desired solution?

(2) How do they behave as we vary the regularization strength that should interpolate between the reference policy and a preference-optimal policy?

(3) On real preference datasets, does EXPO provide practically meaningful gains, and what are their risks and limitations?

Throughout, the competing methods are DPO, IPO, a representative f-DPO variant using the Jensen-Shannon divergence (JS-DPO), and the proposed EXPO losses in their compositional and regression forms.

4.1 Synthetic bandit experiments

The first synthetic experiment adopts a simple bandit setting with a single prompt and three candidate responses/actions, $\mathcal{Y} = \{a_1, a_2, a_3\}$. Human preferences are constructed via a BT-optimal policy $\pi_{\text{BT}}(\cdot)$: two actions, say a_1 and a_2, have similar but slightly different preference

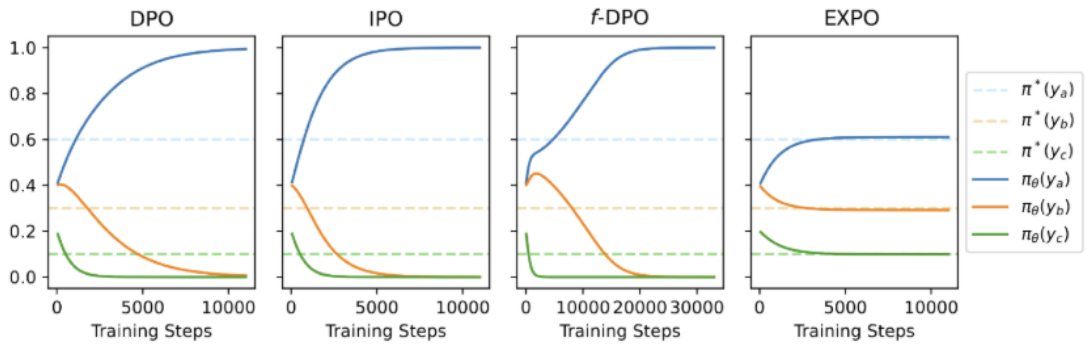

Figure 1: Synthetic 3-arm bandit experiment with small regularization λ. Dashed lines indicate the Bradley-Terry (BT) optimal probabilities for each action, while solid lines show the learned policy over training iterations. EXPO converges to the full BT-optimal distribution, preserving the relative probabilities of near-optimal actions, whereas QPO-type objectives collapse toward a single "best" arm, losing generative diversity.

probabilities, while a_3 is clearly disfavored. From this ground-truth policy, the authors sample pairwise comparisons to form a preference dataset $\mathcal{D}$, and then train different preference optimization objectives starting from the same initialization and reference policy.

Figure 1 plots the learning curves of the predicted action probabilities for each method at a small regularization strength λ. EXPO converges to the dashed BT-optimal lines, preserving the relative probabilities of a_1 and a_2 while appropriately down-weighting a_3. In contrast, DPO, IPO, and JS-DPO all converge to a degenerate solution that concentrates all probability mass on the single most preferred action, consistent with the "weak interpolation" behavior analyzed in the theory. In other words, they behave like a classical RL agent that maximizes reward but ignores the richer structure of the preference distribution.

Figure 2 shows, for each method, the final probability assigned to each action as λ moves from "preference-dominated" (small) to "reference-dominated" (large). Ideally, we would like the family $\{\pi_\theta^\lambda\}$ to smoothly interpolate between π_{BT} and π_{ref}.

The plots confirm the theoretical findings of [10]: EXPO tracks the BT-optimal distribution for small λ, gradually morphs into the reference distribution as λ increases, and never collapses into a degenerate delta on a single action. The QPO family instead interpolates between an extremum and the reference policy, thus satisfying at best the weaker interpolation criterion. From an RL standpoint, EXPO behaves more like a principled regularized policy optimization method, where the regularization parameter interpolates between two meaningful endpoints.

The second synthetic experiment introduces two prompts $x \in \{x_{\mathrm{good}}, x_{\mathrm{bad}}\}$, sampled with equal probability. For x_{good}, the reference policy π_{ref} is already close to the BT-optimal policy π_{BT}, whereas for x_{bad} it is far from optimal. This reflects a realistic alignment scenario: pre-trained LLMs often perform well on many inputs but poorly on specific subsets (e.g., safety-critical or reasoning-heavy prompts). Folloiwng [10], the learning objective should ideally:

(1) preserve π_{ref} where it is already good (on x_{good}); and

(2) move the policy towards π_{BT} where the reference is bad (on x_{bad}).

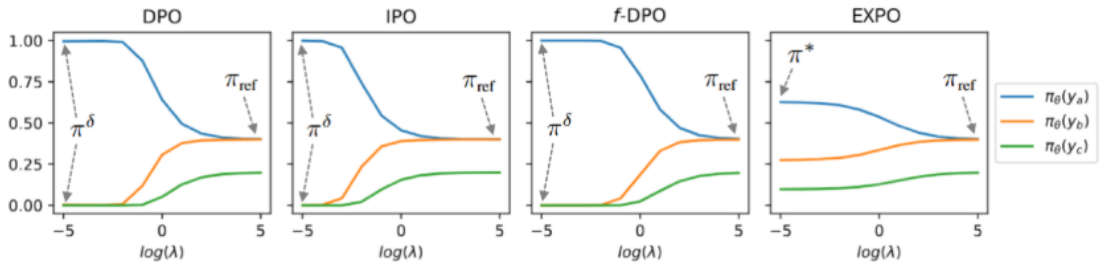

Figure 2: Converged policies in the synthetic bandit as a function of the regularization strength λ. For small λ, EXPO matches the BT-optimal policy, and as λ increases it smoothly returns to the reference policy π_{ref}. In contrast, DPO and related QPO methods interpolate between a mode-collapsed delta policy and the reference, never recovering the full BT-optimal distribution in the low-regularization limit.

Figure 3 plots, for varying λ, a measure of distance between the learned policy π_θ and the BT-optimal policy π_{BT}, separately for x_{good} (top) and x_{bad} (bottom). For QPO methods, as λ is reduced to prioritize alignment with human preferences, the policy indeed moves towards π_{BT} on the bad prompts; however, the same uniform regularization also pulls the policy away from π_{BT} on the good prompts, degrading performance where the reference model was already strong. This behavior matches the "failure to preserve optimal policies" proved in Theorem 3.1. In contrast, EXPO improves the policy on bad prompts while leaving good prompts essentially unchanged across a broad range of λ, enabling targeted improvement without sacrificing behavior that was already aligned.

This experiment highlights both a potential advantage and a risk of RL-inspired alignment. EXPO's explicit design makes the behavior of the regularization parameter more interpretable and controllable. There is a clear separation between regions where policy changes are desirable and regions where they are not. In contrast, implicit-reward methods tied to QPO losses offer less control, as tuning λ trades off improvements in some parts of the state space against degradation elsewhere, which may be problematic in safety-sensitive applications.

4.2 Real-world preference alignment

The final set of experiments moves from synthetic bandits to real-world preference datasets. The authors fine-tune a Pythia 2.8B model [2] on two benchmarks:

- **Anthropic HH** [1, 6]: a large dataset of helpfulness and harmlessness preferences over assistant responses.
- **IMDb** [11, 19]: sentiment-focused preferences where one response is judged better than another for a given movie-review style prompt.

The pipeline follows a common RLHF-style setup: first, a supervised-fine-tuning phase trains the model to imitate high-quality reference responses; then, preference optimization is performed starting from this model as π_{ref}. All methods are trained offline using the same preference tuples,

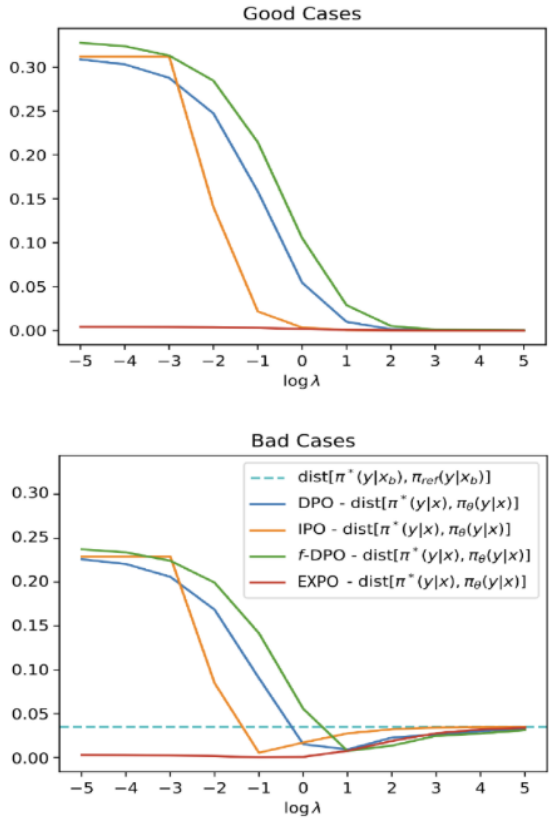

Figure 3: Preservation experiment with two prompt types. The top panel shows the distance between the learned policy and the BT-optimal policy on "good" prompts where the reference policy is already near-optimal; the bottom panel shows the same distance on "bad" prompts where the reference is far from optimal. EXPO preserves good behavior across a wide range of regularization strengths while still improving on bad prompts, wheras QPO objectives improve the bad region at the cost of degrading performance where the base model was already aligned.

and the final policies are evaluated using GPT-4 as a judge, which compares the generated responses and outputs a win rate.

Figure 4 reports GPT-4 win rates on Anthropic HH and IMDb for SFT, DPO, IPO, and the two EXPO variants. On both datasets, EXPO substantially improves over DPO and IPO in terms of win rate, even though all methods start from the same reference and use the same underlying data. Hence, EXPO's explicit treatment of preferences and regularization appears to match human judgment better than purely implicit objectives.

From these numerical experiments, we conclude that EXPO maintains the RL intuition of balancing reward vs. deviation from a reference policy while discarding the heavy RL machinery (explicit reward model, PPO). This yields a more stable and easier-to-implement training loop, which is a major practical advantage.

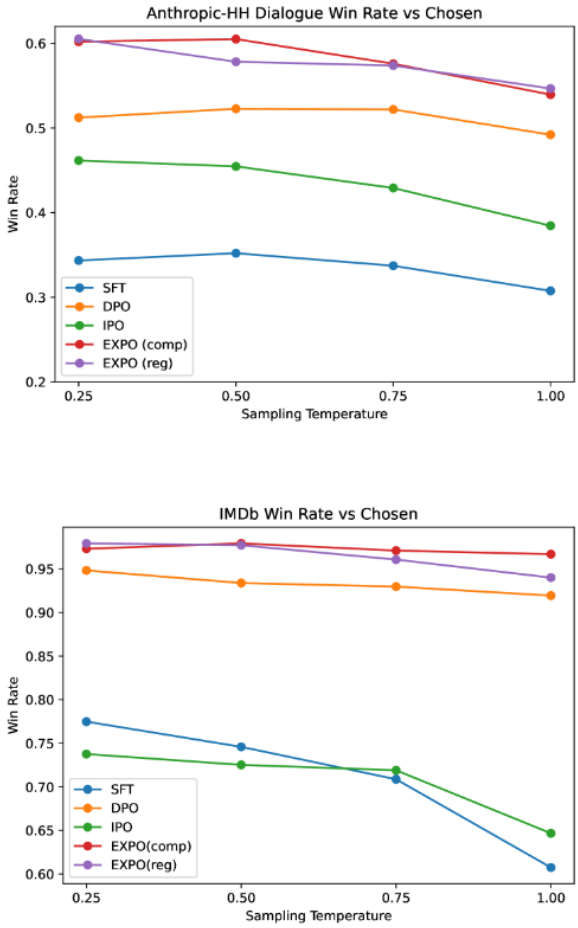

Figure 4: GPT-4 win rates of different preference-optimization methods on the Anthropic HH and IMDb preference datasets, starting from the same supervised-fine-tuned reference model. DPO and IPO improve over SFT, but the EXPO variants achieve higher win rates while maintaining more desirable interpolation and preservation properties, illustrating that explicit preference optimization can be both principled and practically competitive.

5 Conclusion

Standard DPO and its variants derive their loss function directly from the analytical solution of the KL-constrained RL objective. However, this report highlights that strictly adhering to this RL derivation forces the model into counter-intuitive behaviors as identified in Section 2. The EXPO framework demonstrates that we do not need to treat alignment strictly as an "RL problem with an implicit reward." Instead, we can frame it as Explicit Preference Optimization, directly optimizing the likelihood of preferred responses and regularizing against the reference model without the constraints of a reward maximization formula.

Unfortunately, the authors focus on the limitations that appear: The experiments are entirely offline and rely on GPT-4 as an evaluator, which introduces potential biases and makes absolute numbers hard to compare across papers. The BT model is an approximation of human preferences, and the synthetic environments are deliberately simple. Nevertheless, taken

together, the numerical results in Section 4 support the central claim of the paper: we can capture much of the benefit of RLHF-style alignment using a carefully designed explicit preference objective, without running a full RL algorithm, and doing so yields both theoretical clarity and empirical gains. Overall, EXPO can be particularly attractive for practical fine-tuning scenarios where learning a randomized policy is desirable, preserving the diversity and nuance of human preferences rather than collapsing to a single, repetitive output.

A Proof of Theorem 3.1 (Failure to preserve optimal policies)

Setup: Assume the prompt space partitions into disjoint sets $\mathcal{X}_{\text{good}}$ and $\mathcal{X}_{\text{bad}}$. On $\mathcal{X}_{\text{good}}$, the reference policy is already optimal ($\pi_{\text{ref}} = \pi^\star$), whereas on $\mathcal{X}_{\text{bad}}$, it is suboptimal.

Proof. The global QPO loss decomposes into contributions from both regions:

$$\mathcal{L}_{\text{QPO}}(\pi_\theta) = \mathbb{E}_{x\in\mathcal{X}_{\text{good}}}[\ell(u(x))] + \mathbb{E}_{x\in\mathcal{X}_{\text{bad}}}[\ell(u(x))],$$

where $u(x)$ represents the shift in log-ratios relative to the reference:

$$u(x) = \log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)} - \log \frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\text{ref}}(y_l|x)}.$$

(1) **Requirement for $\mathcal{X}_{\text{bad}}$:** To improve performance on bad prompts, the optimizer must shift the policy towards the preference y_w, requiring $u(x) \neq 0$ (specifically, moving in the direction of preference).

(2) **Requirement for $\mathcal{X}_{\text{good}}$:** Since $\pi_{\text{ref}} = \pi^\star$ on these prompts, the optimal shift is exactly $u(x) = 0$. Any deviation $u(x) \neq 0$ moves π_θ away from $\pi^\star$.

(3) **The Conflict:** QPO losses apply *uniform regularization* via the convex function ℓ. The gradient updates required to minimize the loss on $\mathcal{X}_{\text{bad}}$ force a non-zero shift in the shared parameters. Because the objective does not decouple the regularization strength based on optimality, the global minimizer finds a compromise where $u(x) \neq 0$ everywhere.

Consequently, improving the policy on $\mathcal{X}_{\text{bad}}$ unavoidably introduces a non-zero shift on $\mathcal{X}_{\text{good}}$, degrading the policy from its originally optimal state $\pi^\star$. □

B Extended Literature Review: Implicit vs Explicit Preference Optimization[1]

The dominant paradigm for aligning large language models with human values is reinforcement learning from human feedback (RLHF). Normally, a supervised-fine-tuned (SFT) model is used to propose candidate responses, a separate reward model is trained on human pairwise preferences using a Bradley–Terry likelihood, and the policy is then optimized with PPO under a KL penalty to the reference model [4, 20, 16, 12, 1, 15]. This is an *explicit* preference optimization approach: human feedback is first distilled into an explicit scalar reward function

[1] "This section serves the purpose of making up for my incapability of presenting the current project." - N.D.

$r_\phi(x, y)$, and this reward is then used as the objective in a standard regularized RL problem. Recent analyses highlight both the strengths and weaknesses of this approach: it provides a clear decision-theoretic story and strong empirical performance, but it is computationally expensive and sensitive to optimization instabilities [14, 17].

A large body of recent work aims to simplify this process via *implicit* preference optimization, where the reward is not modeled as a separate network but is instead encoded in the policy itself. Direct Preference Optimization (DPO) [13] is the stat-of-the-art method. Starting from the RLHF objective, DPO observes that if the optimal regularized policy can be written in closed form in terms of a reward, then the reward can be reparameterized as a log-ratio between the learned policy and the reference policy. Plugging this implicit reward into the Bradley–Terry model yields a simple logistic loss on log-odds differences that can be optimized offline via stochastic gradient descent, with no separate reward model and no PPO loop. Identity Preference Optimization (IPO) [7] modifies the link function used in DPO's derivation to obtain a loss that is argued to be more robust in near-deterministic preference regimes. Generalized Preference Optimization (GPO) [18] shows that DPO, IPO and related methods can be expressed in a unified form using a convex (or quasi-convex) shaping function applied to log-odds differences, while f-DPO [19] replaces the reverse KL penalty in DPO with general f-divergences, further expanding the design space. ORPO [9] goes one step further by removing the explicit reference model and adding an odds-ratio penalty directly on the SFT likelihood, but it remains implicit in the sense that the preference signal is injected through the policy's own log-probabilities rather than a standalone reward network. Empirical studies such as Devanathan et al. [5] and Cho et al. [3] examine how these implicit objectives behave under different data-collection schemes and label granularities, revealing that while they are easy to train and often competitive with RLHF, their behavior can be sensitive to modeling choices and data biases.

The EXPO framework [10] can be viewed as a return to explicit preference optimization, but within the computationally convenient offline setting. Rather than encoding preferences implicitly via a log-ratio, EXPO starts from a clear notion of optimality—the Bradley–Terry-optimal policy that matches the underlying preference distribution—and builds losses that explicitly involve KL divergences between (i) the BT preference distribution and the preference distribution induced by the policy, and (ii) the learned policy and the reference policy. Hu et al. [10] show that a broad quasi-convex family of implicit objectives (the QPO family, which includes DPO, IPO, GPO and f-DPO) cannot, in general, preserve already-optimal behavior or interpolate smoothly between the reference policy and the BT-optimal policy as regularization is varied, whereas their explicit EXPO losses are constructed to satisfy these properties by design. In this sense, EXPO sits between classical RLHF and implicit methods, as it retains the explicit structure, but avoids the overhead of training a separate reward model and running online RL, making it an appealing test case for understanding when RL concepts are helpful even if ones does not run a full RL algorithm in practice.

References

[1] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[2] S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. van der Wal. Pythia:

A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[3] J. H. Cho et al. VPO: Leveraging the number of votes in preference optimization. *arXiv preprint arXiv:2410.22891*, 2024.

[4] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, pages 4299–4307, 2017.

[5] R. Devanathan, V. Nathan, and A. Kumar. The paradox of preference: A study on LLM alignment algorithms and data acquisition methods. In *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*, 2024.

[6] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

[7] M. Gheshlaghi Azar and coauthors. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2403.00000*, 2024.

[8] H. J. Greenberg and W. P. Pierskalla. A review of quasi-convex functions. *Operations Research*, 19(7):1553–1570, 1971.

[9] J. Hong, Z. Lin, et al. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

[10] S. Hu et al. Explicit preference optimization: No need for an implicit reward model. *arXiv preprint arXiv:2506.07492*, 2025.

[11] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, 2011.

[12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[13] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[14] R. Ramamurthy, P. Ammanabrolu, K. Brantley, J. Hessel, R. Sifa, C. Bauckhage, H. Hajishirzi, and Y. Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. In *International Conference on Learning Representations*, 2023.

[15] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. In *Proceedings of the 34th International Conference on Machine Learning (ICML) Deep RL Workshop*, 2017.

[16] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021, 2020.

[17] G. Swamy et al. All roads lead to RLHF: A unified view of preference optimization methods. *arXiv preprint arXiv:2501.00000*, 2025.

[18] Y. Tang, Z. D. Guo, Z. Zheng, D. Calandriello, R. Munos, M. Rowland, P. H. Richemond, M. Valko, B. Avila Pires, and B. Piot. Generalized preference optimization: A unified approach to offline alignment. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[19] C. Wang, Y. Jiang, C. Yang, H. Liu, and Y. Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *International Conference on Learning Representations*, 2024.

[20] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. F. Christiano, and G. Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.