Back to Blackwell: Closing the Loop on Intransitivity in Multi-Objective Preference Fine-Tuning

Jiahao Zhang[*]
Carnegie Mellon University
Pittsburgh, PA 15213
jiahaozhang@cmu.edu

Lujing Zhang[*]
Carnegie Mellon University
Pittsburgh, PA 15213
lujingz@andrew.cmu.edu

Keltin Grimes
Carnegie Mellon University
Pittsburgh, PA 15213
kgrimes@andrew.cmu.edu

Zhuohao Yu
Carnegie Mellon University
Pittsburgh, PA 15213
zhuohaoy@andrew.cmu.edu

Gokul Swamy
Carnegie Mellon University
Pittsburgh, PA 15213
gswamy@cmu.edu

Zhiwei Steven Wu
Carnegie Mellon University
Pittsburgh, PA 15213
zstevenwu@cmu.edu

Abstract

A recurring challenge in preference fine-tuning (PFT) is handling *intransitive* (i.e., cyclic) preferences. Intransitive preferences often stem from either *(i)* inconsistent rankings along a single objective or *(ii)* scalarizing multiple objectives into a single metric. Regardless of their source, the downstream implication of intransitive preferences is the same: there is no well-defined optimal policy, breaking a core assumption of the standard PFT pipeline. In response, we propose a novel, game-theoretic solution concept, the *Maximum Entropy Blackwell Winner* (*MaxEntBW*), that is well-defined under multi-objective intransitive preferences. To enable computing MaxEntBWs at scale, we derive PROSPER: a provably efficient PFT algorithm. Unlike prior self-play techniques, PROSPER directly handles multiple objectives without requiring scalarization. We then apply PROSPER to the problem of fine-tuning large language models (LLMs) from multi-objective LLM-as-a-Judge feedback (e.g., rubric-based judges), a setting where both sources of intransitivity arise. We find that PROSPER outperforms all baselines considered across both instruction following and general chat benchmarks.

1 Introduction

Author Ted Chiang famously described large language models (LLMs) as a "*blurry JPEG of the web*" [Chiang, 2023]. One way of interpreting his words is that after being trained on vast swathes of internet text, the choices of an LLM "judge" [Zheng et al., 2023] are often inconsistent. Thus, when LLMs are used to rank a set of options $\{A, B, C\}$, they often exhibit *intransitive* (i.e., cyclic) preferences where they rank $A \succ B$ and $B \succ C$, but $A \not\succ C$ [Xu et al., 2025].

By construction, there is no *total ordering* over options under intransitive preferences, as *every* option loses to another. Thus, the standard *preference fine-tuning* (PFT) pipeline [Stiennon et al., 2020, Ouyang et al., 2022, Song et al., 2024, Swamy et al., 2026], in which one attempts to learn such a total ordering via training a reward model, is fundamentally unequipped to robustly handle the intransitive preferences we see when attempting to learn from noisy LLM judge feedback [Procaccia and Rosenschein, 2006, Boutilier et al., 2012, Sorensen et al., 2024, Gölz et al., 2025].

In response, prior work has primarily explored two mitigatory strategies. First, to reduce the baseline inconsistency of LLM judge feedback, recent work has proposed to ask the LLM judge

[*]Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

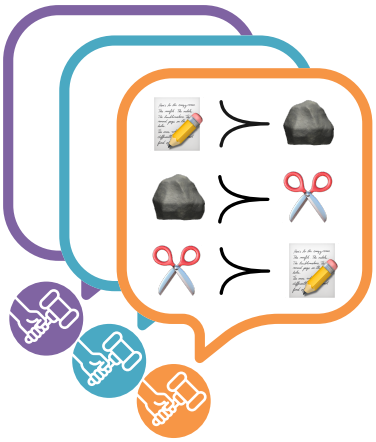


Figure 1: We study the problem of learning from multiple preferences over different objectives or criteria, each of which might be *intransitive* (i.e., inconsistent). Such preferences are common when learning from large language model (LLM) "judges" [Zheng et al., 2023] that evaluate a response along multiple objectives. We propose a novel solution concept for multi-object preference fine-tuning (PFT), the *Maximum Entropy Blackwell Winner (MaxEntBW)*, that remains well-defined under multiple intransitive preferences. We then derive PROSPER: a regression-based algorithm for computing MaxEntBWs, before using it to fine-tune LLMs from multi-objective LLM judge feedback on multiple problems.

to explicitly evaluate candidate responses along multiple objectives, rather than making a single aggregate judgment [Viswanathan et al., 2025, Gunjal et al., 2025]. This echoes observations from psychology, where one of the most common root causes of intransitivity is *scalarization* of multiple objectives into a single metric [Tversky, 1969].

Second, to tackle the remaining per-objective inconsistency, a variety of authors have proposed to frame PFT as zero-sum game solving [Munos et al., 2023, Swamy et al., 2024, Rosset et al., 2024], building on ideas from social choice theory on how to robustly handle intransitive preferences [Kreweras, 1965, Fishburn, 1984]. Intuitively, these approaches attempt to find a policy that the LLM judge prefers as often as possible against the worst-case comparator – a well-defined notion even under intransitive preferences.

Given the complementary nature of these two lines of work, one may naturally hope to combine them into a single, gestalt solution. However doing so is technically challenging for a rather subtle reason: when one splits an aggregate judgement into multiple objectives, the Minimax Theorem [von Neumann, 1928] no longer holds, as famously observed by Blackwell [1956]. The downstream implication of this fact is that standard game-solving machinery (e.g., in the proofs of Swamy et al. [2024]) breaks down, making it unclear how to directly extend prior algorithms that build on said machinery to the multi-objective PFT setting.

Faced with this roadblock, we circle back and ask a first-principles question about the multi-objective PFT setting: *what should it mean to do well when evaluated under multiple objectives, each of which may be globally inconsistent?*

Our answer is the ***Maximum Entropy Blackwell Winner (MaxEntBW):*** *a policy that is preferred to all other policies within its local neighborhood, regardless of the criteria it is judged under.* Remarkably, even though it is a strictly more general solution concept than those explored in prior work, the underlying structure of MaxEntBWs means one can efficiently compute them without requiring the Minimax Theorem to hold. Although MaxEntBWs are defined via a game between two LLM players, their structure permits an efficient reduction to ***a single-player optimization*** problem: only the learner's strategy must be maintained and updated, and the update can be implemented via square-loss regression. Thus, we can capitalize on the more accurate feedback we get from LLM judges by asking them to explicitly evaluate responses along multiple objectives without needing to complicate the policy optimization pipeline. We call our algorithm PROSPER, for **PR**eference-based **O**ptimization with a **S**ingle **P**layer over **E**ntire **R**ubrics. Explicitly, our contributions are threefold:

1. We propose the Maximum Entropy Blackwell Winner (MaxEntBW) as a novel solution concept for multi-objective preference fine-tuning. In contrast to classical solution concepts like the Minimax Winner that are only catered to single objectives [Kreweras, 1965, Fishburn, 1984, Dudík et al., 2015], MaxEntBWs are well-defined under multiple objectives. MaxEntBWs also generalize the Blackwell Winners proposed by Bhatia et al. [2020].

2. We derive PROSPER, a provably efficient, regression-based algorithm for computing MaxEntBWs. We leverage the underlying structure of MaxEntBWs to derive a simple and scalable algorithm for computing them. Doing so requires fusing insights from Ziebart et al. [2008] and Gao

et al. [2024] to first eliminate the inner-loop optimization, before reducing the outer-loop optimization down to simple regression. We prove rigorous performance and efficiency guarantees for PROSPER.

3. We show that multi-objective LLM judge feedback is substantially intransitive in practice, and that PROSPER leads to strong downstream policies in this setting. We first empirically verify that splitting up objectives reduces but does not eliminate intransitivity. We then apply PROSPER to instruction following and general chat and find that it outperforms all baselines considered.

Due to space constraints, we defer the discussion of related work to appendix A.

2 Back to Blackwell: A New Solution Concept for Multi-Objective PFT

2.1 Preliminaries

We begin by discussing relevant background information.

Notation. For any integer n, let $[n]$ denote the set $\{1, \cdots, n\}$. Let $\mathcal{X}$ be the space of prompts and $\mathcal{Y}$ the space of responses. We consider a class of language models (policies) $\Pi \subseteq \{\mathcal{X} \to \Delta(\mathcal{Y})\}$, which we assume is convex and compact. For every prompt $x \in \mathcal{X}$, a policy $\pi \in \Pi$ induces a distribution over responses $\pi(x) \in \Delta(Y)$.

We assume access to a *multi-objective pairwise judge*

$$\mathcal{P} : \mathcal{Y} \times \mathcal{Y} \times \mathcal{X} \to [0,1]^{m(x)},$$

which compares responses $y, y' \in \mathcal{Y} \times \mathcal{Y}$ along $m(x)$ objectives, returning a vector of size $m(x)$. Rather than assuming a fixed set of criteria for every prompt (e.g., helpfulness, harmlessness), this formulation accounts for the diversity of possible prompts fed to today's LLMs: each prompt may be associated with different (and different numbers of) relevant criteria. For each $k \in [m(x)]$, $\mathcal{P}^k(y \succ y' \mid x)$ denotes the probability that $y \succ y'$, conditioned on prompt x and based on criterion k. We emphasize that we make no assumptions about the transitivity of preferences along each objective. With a slight abuse of notation, we can extend this to a preference function between two policies $\pi, \pi' \in \Pi \times \Pi$: $\mathcal{P}(\pi \succ \pi' \mid x) \triangleq \mathbb{E}_{y \sim \pi(x),\, y' \sim \pi'(x)}[\mathcal{P}(y \succ y' \mid x)] \in [0,1]^{m(x)}$.

von Neumann Winner. Under intransitive preferences, there is often no well-defined optimal policy (more formally, no *Condorcet Winner*, Brandt et al. [2016]), as every response y loses to some other response y'. Thus, we need to consider alternative solution concepts. In the single-objective setting, a popular choice is the *Maximal Lottery* [Kreweras, 1965, Fishburn, 1984] or *von Neumann Winner* [Dudík et al., 2015], which is the policy that is as preferred as possible against its worst-case opponent. More formally, a von Neumann Winner is a Nash Equilibrium of the zero-sum game with $\mathcal{P}$ as the payoff matrix. Due to the symmetry of the payoff matrix, simple self-play is provably efficient for computing such equilibria [Swamy et al., 2024].

Blackwell Winner. A *Blackwell Winner* Bhatia et al. [2020] is the natural extension of the von Neumann Winner to the multi-objective setting, drawing on Blackwell's notion of a *target set* [Blackwell, 1956]. For simplicity, we present this solution concept in the promptless setting. For vector z and set C, define the ℓ_∞ *distance function* as $\text{dist}_\infty(z, C) = \min_{c \in C} ||z - c||_\infty$.

Suppose we want our policy π to be preferred with at least probability p to any policy $\pi' \in \Pi$ along all m objectives: $\forall \pi' \in \Pi, \mathcal{P}(\pi \succ \pi') \in [p, \infty]^m$, where $C(p) \triangleq [p, \infty]^{m(x)}$ is the corresponding target set. For any $p \geq 1/2$, we can *linearize* the distance function as

$$\max_{\pi' \in \Pi} \text{dist}_\infty(\mathcal{P}(\pi \succ \pi'), C(p)) = \max_{\pi' \in \Pi} \left(p - \min_{w \in \Delta_m} \langle w, \mathcal{P}(\pi \succ \pi') \rangle\right),$$

since π' can ensure that any coordinate of $\mathcal{P}^k(\pi \succ \pi') \leq 1/2$ (e.g., by setting $\pi' = \pi$). In this case, we can define the ℓ_∞ Blackwell Winner [Bhatia et al., 2020] as

$$\arg\min_{\pi \in \Pi} \max_{\pi' \in \Pi} \left(p - \min_{w \in \Delta_m} \langle w, \mathcal{P}(\pi \succ \pi') \rangle\right). \tag{1}$$

2.2 The Maximum Entropy Blackwell Winner

Note that removing the constant shift p in Equation (1) does not affect the solution. Then, reintroducing the prompts x, we can express the ℓ_∞ Blackwell Winner as the solution to

$$\max_{\pi \in \Pi} \min_{w:\mathcal{X} \to \Delta_{m(x)}} \min_{\pi' \in \Pi} \mathbb{E}_x[\langle w(x), \mathcal{P}(\pi \succ \pi' \mid x)\rangle]. \tag{2}$$

While perhaps reasonable for small-scale problems, performing a *best response* over adversary policies π' as in the above is infeasible at the scale of today's foundation models. Furthermore, adversarially selecting both the criteria and comparator policy could lead to situations where *all* policies look bad, even though some are qualitatively better than others. Thus, to control the strength of the adversary, we propose to KL regularize it to a reference policy π_{ref}:

$$\max_{\pi \in \Pi} \min_{w:\mathcal{X} \to \Delta_{m(x)}} \min_{\pi' \in \Pi} \mathbb{E}_x[\langle w(x), \mathcal{P}(\pi \succ \pi' \mid x)\rangle + \beta\, \mathbb{D}_{\mathrm{KL}}(\pi'(x) \,\|\, \pi_{\mathsf{ref}}(x))], \tag{3}$$

where $\beta > 0$ controls the strength of the regularization. Beyond being perhaps the natural choice, the use of KL regularization on the adversary will allow us to elide adversarial training, as we discuss in greater detail below. For now though, let us define the game value of a policy π as

$$V(\pi) \triangleq \min_{w:\mathcal{X} \to \Delta_{m(x)}} \min_{\pi' \in \Pi} \mathbb{E}_x\big[\langle w(x), \mathcal{P}(\pi \succ \pi' \mid x)\rangle + \beta\, \mathbb{D}_{\mathrm{KL}}(\pi'(x) \,\|\, \pi_{\mathsf{ref}}(x))\big]. \tag{4}$$

We are now equipped to define the *Maximum Entropy Blackwell Winner (MaxEntBW)*:

Definition 2.1 (Maximum Entropy Blackwell Winner (MaxEntBW))**.** A policy $\pi^\star \in \arg\max_{\pi \in \Pi} V(\pi)$ is called a *Maximum Entropy Blackwell Winner (MaxEntBW)*.

In words, a MaxEntBW is a policy that compares favorably to policies in its local neighborhood, regardless of the objective it is evaluated under – a nuanced form of robustness. Observe that this solution concept is well-defined even under multiple, globally inconsistent sets of preferences.

In contrast to the single-objective setting, the Minimax Theorem [von Neumann, 1928] doesn't hold in the multi-objective setting when we consider the worst case over objectives [Blackwell, 1956]. Thus, the *order of play* matters: the value of the game is different depending on whether learner π or adversary π' moves first [Roth, 2023]. This means that the self-play approaches [Swamy et al., 2024] that were sufficient in the single-objective setting cannot be directly extended to the multi-objective setting, seemingly indicating that adversarial training may be necessary.

3 PROSPER: PReference Optimization with a Single Player over Entire Rubrics

Despite the apparent complexity of solving Equation (3) to compute a MaxEntBW (Definition 2.1), the underlying structure of the optimization problem admits a strikingly simple solution, as we now explore. We proceed by working from the inner-most optimization problem outwards.

Key Step #1: Entropy Regularization Ensures a Closed-Solution for π'. Fix a π and w. Then, we need to solve

$$\min_{\pi' \in \Pi} \mathbb{E}_x[\langle w(x), \mathcal{P}(\pi \succ \pi' \mid x)\rangle + \beta\, \mathbb{D}_{\mathrm{KL}}(\pi'(x) \,\|\, \pi_{\mathsf{ref}}(x))]. \tag{5}$$

Using $\mathcal{P}_\pi(y' \mid x) = \mathbb{E}_{y \sim \pi(x)}[\mathcal{P}(y \succ y' \mid x)]$, we know from Ziebart et al. [2008] that this problem has solution

$$\pi'_\star(y' \mid x) = \frac{\pi_{\mathsf{ref}}(y' \mid x) \exp(-\langle w(x), \mathcal{P}_\pi(y' \mid x)\rangle/\beta)}{Z(w, \pi \mid x)}, \tag{6}$$

where $Z(w, \pi \mid x) = \mathbb{E}_{y' \sim \pi_{\mathsf{ref}}(x)}[\exp(-\langle w(x), \mathcal{P}_\pi(y' \mid x)\rangle/\beta)]$ is the *partition function* that ensures the above sums to 1 across $\mathcal{Y}$. As in Gupta et al. [2025], we can then plug this closed-form expression back into Equation (3) to arrive at

$$\max_{\pi \in \Pi} \min_{w:\mathcal{X} \to \Delta_{m(x)}} \mathbb{E}_x[-\beta \log Z(w, \pi|x)]. \tag{7}$$

Observe that this optimization problem does not require adversarial training between policies to solve.

Algorithm 1 PROSPER: PReference Optimization with a Single Player over Entire Rubrics

Input: A dataset of prompts $\mathcal{D}$, reference policy π_{ref}, judge $\mathcal{P} : \mathcal{Y} \times \mathcal{Y} \times \mathcal{X} \to [0,1]^{m(x)}$, KL penalty coefficient β, time steps T, step size η, policy class $\Pi = \{\pi_\theta : \theta \in \Theta\}$, sample size M for gradient estimation.

Initialize: $\pi_{\theta_0} = \pi_{\text{ref}}$.

for $t = 0, \ldots, T-1$ **do**

For each $x \in \mathcal{D}$, collect $\mathcal{D}(x) = \{(y_1, ..., y_M, y'_1, ..., y'_M, z, z')\}$, where $y_i \sim \pi_{\theta_t}(x)$, $y'_j \sim \pi_{\text{ref}}(x)$, $z \sim \pi_{\theta_t}(x)$, $z' \sim \pi_{\text{ref}}(x)$.

For each $x \in \mathcal{D}$, compute the worst-case objective $\widehat{k}(x)$ via Eq. 12.

For each $x \in \mathcal{D}$, compute gradient estimates $\widehat{g}_t(x, z)$ and $\widehat{g}_t(x, z')$, where

$$\widehat{g}_t(x,z) = \frac{\sum_{j=1}^{M} \mathcal{P}^{\widehat{k}(x)}(z \succ y'_j \mid x) \exp\left(-\frac{1}{M\beta} \sum_{i=1}^{M} \mathcal{P}^{\widehat{k}(x)}(y_i \succ y'_j \mid x)\right)}{\sum_{j=1}^{M} \exp\left(-\frac{1}{M\beta} \sum_{i=1}^{M} \mathcal{P}^{\widehat{k}(x)}(y_i \succ y'_j \mid x)\right)}.$$

Update policy parameters via solving a square loss regression problem:

$$\theta_{t+1} = \arg\min_{\theta \in \Theta} \sum_{\substack{x \in \mathcal{D} \\ z, z' \in \mathcal{D}(x)}} \left(\frac{1}{\eta} \left(\ln \frac{\pi_\theta(z \mid x)}{\pi_{\theta_t}(z \mid x)} - \ln \frac{\pi_\theta(z' \mid x)}{\pi_{\theta_t}(z' \mid x)} \right) - (\widehat{g}_t(x,z) - \widehat{g}_t(x,z')) \right)^2.$$

end for

Key Step #2: Simplifying w down to a Prompt-Wise Minimum. Next, we observe that Equation (7) is *concave* in $w(x)$. Thus, for $\forall x \in \mathcal{X}$, we know that $w^\star(x)$ must be a vertex of the simplex $\Delta_{m(x)}$ (i.e., a standard basis vector). As the set of objectives is usually relatively small in practice, we can simply take a minimum prompt-wise across different objectives to eliminate the optimization over w. More formally, we can define the kth partition function

$$Z^k(\pi \mid x) = \mathbb{E}_{y' \sim \pi_{\text{ref}}(x)} \left[\exp\Big(- \mathcal{P}^k_\pi(y' \mid x)/\beta \Big) \right], \tag{8}$$

and the worst-case coordinate as $k^\star(x) = \arg\min_{k \in [m(x)]} -\beta \log Z^k(\pi \mid x)$. Then, we can further simplify down Equation (7) to:

$$\max_{\pi \in \Pi} \mathbb{E}_x \Big[-\beta \log Z^{k^\star(x)}(\pi|x) \Big]. \tag{9}$$

We have now reduced our optimization problem, which initially had three players, to a single player.

Key Step #3: Using Regression to Approximate Online Mirror Descent. We now observe that Equation (9) is concave in π. This is because each $-\beta \log Z^k(\pi \mid x)$ is concave in π (see Lemma C.1 in Appendix C) and the minimum of concave functions is also concave. We can therefore reduce this problem to online convex optimization [Hazan, 2023] and use Online Mirror Descent (in particular, with KL as the Bregman Divergence) for policy optimization.

The first step of a descent algorithm is calculating the gradient. Algebra tells us that the (functional) gradient with respect to π at some response $z \in \mathcal{Y}$ of Equation (9) is

$$g(z \mid \pi, x) = \frac{\mathbb{E}_{y' \sim \pi_{\text{ref}}(x)}[\mathcal{P}^{k^\star(x)}(z \succ y' \mid x) \exp(-\mathcal{P}^{k^\star(x)}_\pi(y' \mid x)/\beta)]}{Z^{k^\star(x)}(\pi|x)}. \tag{10}$$

On small-scale problems, we could follow this gradient by performing mirror descent independently at each prompt $x \in \mathcal{X}$ and response $y \in \mathcal{Y}$. Unfortunately, both $\mathcal{X}$ and $\mathcal{Y}$ are too large in modern foundation model training for doing so to be practically feasible. In response, we use the regression-based reformulation of mirror descent proposed by Gao et al. [2024, 2025], Brantley et al. [2025], Zhou et al. [2024a] to arrive at the following update:

$$\theta_{t+1} = \arg\min_{\theta \in \Theta} \mathbb{E}_{x, z \sim \pi_{\theta_t}(x),\, z' \sim \pi_{\text{ref}}(x)} \left(\frac{1}{\eta} \left(\ln \frac{\pi_\theta(z \mid x)}{\pi_{\theta_t}(z \mid x)} - \ln \frac{\pi_\theta(z' \mid x)}{\pi_{\theta_t}(z' \mid x)} \right) - \big(g(z \mid \pi_{\theta_t}, x) - g(z' \mid \pi_{\theta_t}, x) \big) \right)^2.$$

As proved by Gao et al. [2024], solving the above optimization problem would correspond to exact mirror descent. While we are unlikely to be able to do this in practice, we can still learn a policy with strong performance guarantees under approximation error, as we show in Section 3.1.

Putting It All Together in the Finite-Sample Setting. In practice, we also need to estimate the infinite-sample gradient (Eq. 10) based on samples from the current policy and the reference policy. Thus, we draw M samples $y_{1:M} \sim \pi(x)$ and another M samples $y'_{1:M} \sim \pi_{\text{ref}}(x)$ independently. We then approximate the partition function

$$\widehat{Z}^k(x) := \frac{1}{M}\sum_{j=1}^{M}\left[\exp\left(-\frac{1}{M\beta}\sum_{i=1}^{M}\mathcal{P}^k(y_i \succ y'_j \mid x)\right)\right], \tag{11}$$

$$\widehat{k}(x) = \arg\min_{k\in[m(x)]} -\beta \log \widehat{Z}^k(x), \tag{12}$$

and arrive at an estimate of the gradient $\widehat{g}_t(x,z)$:

$$\frac{\sum_{j=1}^{M}\mathcal{P}^{\widehat{k}(x)}(z \succ y'_j \mid x)\exp(-\frac{1}{M\beta}\sum_{i=1}^{M}\mathcal{P}^{\widehat{k}(x)}(y_i \succ y'_j \mid x))}{\sum_{j=1}^{M}\exp(-\frac{1}{M\beta}\sum_{i=1}^{M}\mathcal{P}^{\widehat{k}(x)}(y_i \succ y'_j \mid x))}. \tag{13}$$

We then regress this gradient estimate in terms of the policy to perform updates as in Gao et al. [2024]. We call this approach PROSPER: PReference Optimization with a Single Player over Entire Rubrics and outline it in Algorithm 1.

3.1 Performance Guarantee for PROSPER

Recall that the game value of a policy π is denoted by $V(\pi)$. Let $V(\pi^\star)$ be the optimal value. We now prove that under the assumption that the regression problem can be solved accurately in-distribution, PROSPER can efficiently learn a policy $\hat{\pi}$ such that $V(\hat{\pi})$ is close to $V(\pi^\star)$. Intuitively, this means $\hat{\pi}$ competes around as favorably as possible with policies in its local neighborhood, regardless of the objective it is judged under. More formally, we assume that

Assumption 3.1. For all $t \in [T]$, we have that for some ϵ,

$$\mathbb{E}_{\substack{x,\\ z\sim\pi_{\theta_t}(x),\\ z'\sim\pi_{\text{ref}}(x)}}\left(\frac{1}{\eta}\left(\ln\frac{\pi_{\theta_{t+1}}(z|x)}{\pi_{\theta_t}(z|x)} - \ln\frac{\pi_{\theta_{t+1}}(z'|x)}{\pi_{\theta_t}(z'|x)}\right) - (g(z \mid \pi_{\theta_t}, x) - g(z' \mid \pi_{\theta_t}, x))\right)^2 \le \epsilon.$$

Intuitively, this assumption says that there is a policy in our policy class that is able to accurately fit the difference of gradients. We believe this is a reasonable assumption for modern deep networks. Next, given some policy π, we denote the *concentrability coefficient* as $C_{\pi_{\text{ref}}\to\pi} = \max_{x,y}\frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)}$ [Kakade and Langford, 2002], which quantifies how well the reference policy π_{ref} covers π. We are now ready to state our main theoretical result:

Theorem 3.2. *Under Assumption 3.1, after T iterations, there exists some $\hat{\pi} \in \{\pi_{\theta_1}, \cdots, \pi_{\theta_T}\}$ such that $V(\pi^\star) - V(\hat{\pi}) \le O(\sqrt{1/T} + \sqrt{C_{\pi_{\text{ref}}\to\pi^\star}\epsilon})$, where $\pi^\star \in \arg\max_{\pi\in\Pi} V(\pi)$ denotes an arbitrary MaxEntBW.*

In words, this theorem says that as long as supervised learning works (i.e., ϵ is small), then PROSPER can efficiently compute an approximate MaxEntBW. We defer the detailed proof to Appendix C and turn our attention to empirically validating our approach on LLM fine-tuning problems.

4 LLM Fine-Tuning Experiments Setup

We now outline the setup used for our LLM experiments. We defer all hyperparameters to App. D.1.

Training Dataset. We train on the WILDCHECKLISTS dataset [Viswanathan et al., 2025]. WILDCHECKLISTS is based on WILDCHAT, a collection of conversations between users and AI language models [Zhao et al., 2024]. As described in Viswanathan et al. [2025], WILDCHECKLISTS was created by using Qwen2.5-72B-Instruct to generate *prompt-specific* rubrics to guide LLM judge evaluation of candidate responses. Note that our framework – which does not assume a fixed set of criteria across all prompts – is able to flexibly accommodate this more nuanced feedback. One can view each checklist item as one of the multiple objectives discussed in our preceding formulation.

Policy and Judge Models. We fine-tune the 3B and 7B parameter Qwen2.5-Instruct models [Qwen et al., 2025] and use Qwen3-14B as our judge model [Yang et al., 2025]. We selected Qwen3-14B due to the relatively high agreement it demonstrated with author judgments across multiple criteria. We include the full prompt templates in Appendix D and use the standard averaging over both input orderings to mitigate the positional bias of pairwise judges [Zheng et al., 2023]. All LLM judges score response pairs along a checklist item using a 5-point Likert scale. We query the judge five times for each response ordering (ten times in total) and average the scores to reduce variance and mitigate positional bias before normalizing to $[0, 1]$.

Benchmarks. We primarily evaluate our trained models on AlpacaEval [Dubois et al., 2025] and Arena-Hard [Li et al., 2025], as these benchmarks are in-domain for the WildChecklists training dataset. Additionally, to ensure our training procedure doesn't cause regression on other capabilities, we also report results on multiple-choice QA and reasoning benchmarks: IFEval [Zhou et al., 2023], MMLU [Hendrycks et al., 2021], ARC [Clark et al., 2018], HellaSwag [Zellers et al., 2019], and TruthfulQA [Lin et al., 2021].

Baselines. Given the similarity of the training data, we primarily evaluate our methods against the RLCF (RL from Checklist Feedback) method of Viswanathan et al. [2025]. For each prompt, candidate response, and checklist item, RLCF generates a score by querying the LLM judge. Another LLM (Qwen2.5-72B-Instruct in our experiments) then generates a set of weights used to combine the per-item scores into a per-response reward. Thus, the primary distinction between our method and RLCF is the use of a more principled, game-theoretic aggregation of per-item scores. Less important distinctions include the use of DPO [Rafailov et al., 2024] on the highest and lowest scoring responses for a prompt rather than gap size-sensitive REBEL [Gao et al., 2024] for policy optimization and our use of *pairwise* judges rather than a judge that evaluates responses in isolation. To ensure a fair comparison, we re-implement RLCF within our optimization pipeline using REBEL as the policy optimizer and report the resulting performance in the main text. We further evaluate the authors' released checkpoints and find their performance to be comparable to our re-implementation (see appendix B). Both, however, are consistently outperformed by our method.

Ablations. In addition to evaluating our full method, PROSPER, we evaluate two simplifications, each of which correspond to dropping one of the inner loop optimizations.

Ablation #1: Single-Objective Optimization. We ablate the importance of explicitly optimizing over multiple objectives by instead asking the LLM judge for a single, aggregate score (i.e., $\forall x \in \mathcal{X}, m(x) = 1$ that reflects all checklist items *jointly*). We refer to this variant as PROSPER-JC.

Ablation #2: Fixed Competitor Policy. Rather than adversarially selecting a competitor policy, we can instead compare our policy to π_{ref}. This corresponds to taking $\beta \to \infty$ or a *variational bound* on our full objective [Gupta et al., 2025]. Thus, we refer to this variant as PROSPER-VB.

Across all variants of our method, we use $M = 2$ samples for gradient estimation and sample 4 responses per prompt, rather than a single pair (z, z') as in Algorithm 1. We also pool pairs across prompts and only train on the response pairs with the largest gaps in gradient vectors, which we find improves the quality of the final learned model. Intuitively, one can view this process as filtering out response pairs where the LLM judge was not confident in its choice.

5 PROSPER Enables Efficient Multi-Objective PFT From LLM Judge Feedback

Our experiments focus on answering four key questions:

1. How much of an issue is LLM judge intransitivity in practice? We compute the fraction of prompts for which LLM judges have intransitive preferences and for which Condorcet Winner exists. We also evaluate how this fraction changes if we decompose a checklist into individual items and ask the judge to evaluate each independently.

2. Is PROSPER able to optimize multi-objective judge preferences? We compute the pairwise win-rate between all trained pairs of trained policies as in Munos et al. [2023] to quantify how well the policies produced by PROSPER optimize nuanced, multi-criteria LLM judge preferences.

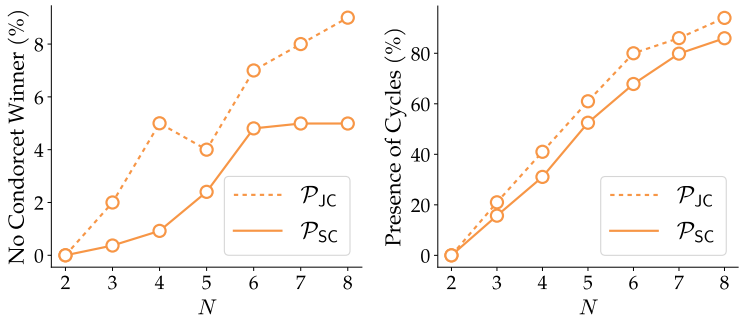


Figure 2: On a held-out set of prompts from WILDCHECKLISTS, we report the fraction of prompts with no Condorcet Winner *(left)* and Intransitive Preferences *(right)* for both the $\mathcal{P}_{\mathsf{JC}}$ (joint check) and $\mathcal{P}_{\mathsf{SC}}$ (single check) judges, where N denotes the number of generated responses. We see that splitting up a rubric into multiple items before passing it to the LLM judge (i.e., using $\mathcal{P}_{\mathsf{SC}}$ rather than $\mathcal{P}_{\mathsf{JC}}$) reduces but doesn't eliminate inconsistent preferences.

3. Do the policies trained via PROSPER exhibit qualitatively strong behavior? Beyond optimizing the MaxEntBW objective, we also evaluate how well the policies produced by PROSPER perform on the ALPACAEVAL (i.e., instruction following), ARENA-HARD (i.e., general chat), IFEVAL, MMLU, ARC, HELLASWAG, and TRUTHFULQA benchmarks. We primarily evaluate on the first 2 benchmarks as they are in-domain for our training set WILDCHECKLISTS, with the rest serving as regression tests.

4. How important are the inner loop optimizations for PROSPER's practical performance? We ablate the effect of the adversarially chosen competitor (PROSPER-VB) and adversarially weighted objectives (PROSPER-JC) on trained policy performance across all seven benchmarks.

We now answer these questions in order.

5.1 Multiple Objectives Mitigate but Don't Fully Eliminate LLM Judge Intransitivity

On a held-out set of 100 prompts from the WILDCHECKLISTS dataset, we evaluate the inconsistency of our LLM judge on N responses from the Qwen2.5-7B-Instruct base model. We compute both the percentage of prompts for which we see intransitive preferences (i.e., those where the pairwise preferences form a cycle of any length [Xu et al., 2025]) and there is no Condorcet Winner (i.e., those where no fixed response is preferred to all others). We note that these two quantities are not the same, as cycles may only occur across relatively weak responses. We compare two judges: $\mathcal{P}_{\mathsf{JC}}$ (joint check) and $\mathcal{P}_{\mathsf{SC}}$ (single check). For $\mathcal{P}_{\mathsf{JC}}$, the LLM judge takes the full checklist as input and outputs a single, aggregate score to indicate how much it prefers one response over the other. For $\mathcal{P}_{\mathsf{SC}}$, the LLM judge evaluates both responses along each checklist item independently. We compute metrics for each (prompt, item) pair before reporting the average over items and prompts.

In Figure 2, we see that both judges frequently have intransitive preferences for large enough batch size N, echoing the findings of Xu et al. [2025]. This underscores the importance of robust algorithms like PROSPER for learning from such inconsistent feedback. We also see that splitting up a checklist into multiple items (i.e., using $\mathcal{P}_{\mathsf{SC}}$ rather than $\mathcal{P}_{\mathsf{JC}}$) improves both metrics but does not completely eliminate the problems of having no Condorcet Winner and intransitive preferences. Thus, while checklist / rubric rewards are an important first step, we still need to take care in the multi-objective setting to handle LLM judge intransitivity.

Model	Arena-Hard		AlpacaEval	
	Vanilla	Style-Controlled	Vanilla	Length-Controlled
Qwen2.5-7B-Instruct	42.4	44.2	37.1	25.32
+ RLCF	42.5	43.9	41.4	17.24
+ PROSPER-JC	44.2	42.0	<u>55.3</u>	**38.21**
+ PROSPER-VB	<u>47.6</u>	<u>45.5</u>	51.2	33.64
+ PROSPER	**49.2**	**46.1**	**55.4**	<u>37.61</u>

Table 1: Highest score is **bolded** and second highest <u>underlined</u>. Across instruction following and general chat benchmarks, PROSPER consistently outperforms the RLCF baseline and PROSPER-JC and PROSPER-VB baselines at the 7B parameter scale.

5.2 PROSPER Effectively Optimizes Multi-Objective LLM Judge Preferences

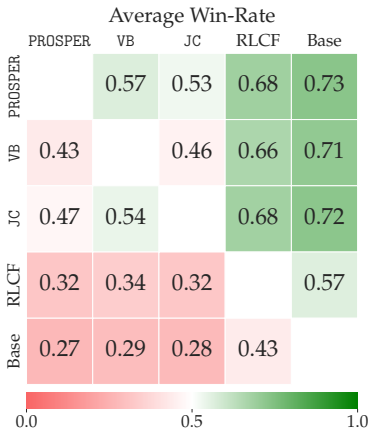


Figure 3: We consistently see policies trained via PROSPER outperform RLCF (roughly $2/3$ of the time), baseline (roughly $3/4$ of the time), and ablation method policies, as measured by LLM judge win-rates on held-out prompts. This indicates that PROSPER is able to more effectively optimize nuanced, multi-criteria LLM judge preferences than the other methods we consider.

Next, we ask the LLM judge to compare all trained 7B policies (PROSPER, PROSPER-JC, PROSPER-VB, RLCF, and the base Qwen2.5-7B-Instruct) on a set of 500 held-out prompts from the test set. For each pair of policies and prompt, we sample responses from each policy. We then ask our judge to evaluate each pair of responses along all criteria and pick a winning policy for each. We then average over checklist items, before finally averaging over prompts to compute a *win rate*.

In Figure 3, we see that PROSPER produces policies that better optimize for judge preferences than the base policy and RLCF baseline. Zooming in on the top row, we see that PROSPER produces policies that generate responses preferred to the base policy responses roughly $3/4$ of the time and the RLCF policy responses roughly $2/3$ of the time. Furthermore, we see that the full version of PROSPER outperforms both the PROSPER-JC and PROSPER-VB ablations, indicating that both inner loop optimizations improve efficacy at optimizing rubric-based LLM judge preferences.

5.3 PROSPER Produces Policies With Strong Performance on Multiple Benchmarks

In Table 1, we report the performance of the 7B policies produced by all methods on AlpacaEval 2.0 (instruction following) and Arena-Hard (general chat). For both benchmarks, we use gpt-5-mini as the LLM judge. We see that PROSPER outperforms RLCF and the PROSPER-JC and PROSPER-VB ablations on Arena-Hard and Vanilla AlpacaEval and is a close second place for length-controlled AlpacaEval. This indicates that beyond merely optimizing judge preferences, PROSPER trains qualitatively strong policies. This also provides evidence that beyond being robust in theory, the MaxEntBW solution concept appears to correspond to qualitatively strong policies.

Furthermore, by comparing the first and last rows of Table 2, we see that PROSPER does not degrade model commonsense reasoning and knowledge QA capabilities (which are out-of-domain for our

Model	IFEval	MMLU	ARC	HellaSwag	TruthfulQA	Avg
Qwen2.5-7B-Instruct	71.16	63.39	85.23	**74.54**	**68.42**	**72.55**
+ RLCF	67.83	**63.47**	85.07	74.18	65.97	71.30
+ PROSPER-JC	69.87	63.37	85.41	74.18	65.61	71.69
+ PROSPER-VB	**73.75**	63.36	85.32	74.04	65.6	72.32
+ PROSPER	69.50	63.42	**85.49**	74.12	65.97	71.70

Table 2: Highest score is **bolded** and second highest underlined. By comparing the first and last row of the table, we see that PROSPER does not lead to significant model commonsense reasoning and general knowledge QA capability degradation at the 7B parameter scale.

training set), indicating that neither our solution concept nor training scheme is so adversarial as to produce overly conservative policies. We find similar results at the 3B scale – see Appendix B.

6 Conclusion

We introduce the *Maximum Entropy Blackwell Winner*: a solution concept for learning from multiple, potentially intransitive preferences. This is frequently the case when learning from rubric-based LLM judges. We then derive PROSPER, a provably efficient algorithm for computing MaxEntBWs, before using it to fine-tune LLMs at the 3B and 7B parameter scales. While our current evaluation focuses on these specific scales and the instruction-following domain, future work may explore scaling PROSPER to larger architectures and more diverse task environments.

Acknowledgments and Disclosure of Funding

We thank Zhaolin Gao, Vijay Viswanathan, and Zhicheng Zhang for their helpful discussions and valuable feedback.

Funding. Zhiwei Steven Wu was supported in part by an NSF CAREER Award and NSF Award No. 2232693.

Competing Interests. The authors declare no competing interests.

Broader Impacts

This paper presents PROSPER, a method for reinforcement learning post-training of large language models. There are many potential societal consequences of improved LLM capabilities and of broader adoption of RL post-training, none of which we feel must be specifically highlighted here.

References

P. A Alamdari, S. Ebadian, and A. D. Procaccia. Policy aggregation. *Advances in Neural Information Processing Systems*, 37:68308–68329, 2024.

A. Agnihotri, R. Jain, D. Ramachandran, and Z. Wen. Multi-objective preference optimization: Improving human alignment of generative models. *arXiv preprint arXiv:2505.10892*, 2025.

M. G. Azar, Z. D. Guo, B. Piot, R. Munos, M. Rowland, M. Valko, and D. Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

H. Bauer. Minimalstellen von funktionen und extremalpunkte. *Archiv der Mathematik*, 9(4):389–393, 1958.

K. Bhatia, A. Pananjady, P. Bartlett, A. Dragan, and M. J. Wainwright. Preference learning along multiple criteria: A game-theoretic perspective. *Advances in neural information processing systems*, 33:7413–7424, 2020.

D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1–8, 1956. doi: 10.2140/pjm.1956.6.1. URL https://projecteuclid.org/journals/pacific-journal-of-mathematics/volume-6/issue-1/An-analog-of-the-minimax-theorem-for-vector-payoffs/pjm/1103044235.

C. Boutilier, I. Caragiannis, S. Haber, T. Lu, A. D. Procaccia, and O. Sheffet. Optimal social choice functions: A utilitarian view. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 197–214, 2012.

F. Brandt, V. Conitzer, U. Endriss, J. Lang, and A. D. Procaccia. *Handbook of computational social choice*. Cambridge University Press, 2016.

K. Brantley, M. Chen, Z. Gao, J. D. Lee, W. Sun, W. Zhan, and X. Zhang. Accelerating rl for llm reasoning with optimal advantage regression. *arXiv preprint arXiv:2505.20686*, 2025.

S. Chakraborty, J. Qiu, H. Yuan, A. Koppel, F. Huang, D. Manocha, A. S. Bedi, and M. Wang. Maxmin-rlhf: Alignment with diverse human preferences. *arXiv preprint arXiv:2402.08925*, 2024.

T. Chiang. Chatgpt is a blurry jpeg of the web. *The New Yorker*, Feb. 2023. URL https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web.

K. Chidambaram, K. V. Seetharaman, and V. Syrgkanis. Direct preference optimization with unobserved preference heterogeneity: The necessity of ternary preferences. *arXiv preprint arXiv:2510.15716*, 2025.

P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018.

V. Conitzer, R. Freedman, J. Heitzig, W. H. Holliday, B. M. Jacobs, N. Lambert, M. Mossé, E. Pacuit, S. Russell, H. Schoelkopf, et al. Social choice should guide ai alignment in dealing with diverse human feedback. *arXiv preprint arXiv:2404.10271*, 2024.

G. Cui, L. Yuan, N. Ding, G. Yao, B. He, W. Zhu, Y. Ni, G. Xie, R. Xie, Y. Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. 2025. URL https://arxiv.org/abs/2404.04475.

M. Dudík, K. Hofmann, R. E. Schapire, A. Slivkins, and M. Zoghi. Contextual dueling bandits. In *Conference on Learning Theory*, pages 563–587. PMLR, 2015.

S. Feng, T. Sorensen, Y. Liu, J. Fisher, C. Y. Park, Y. Choi, and Y. Tsvetkov. Modular pluralism: Pluralistic alignment via multi-llm collaboration. *arXiv preprint arXiv:2406.15951*, 2024.

P. C. Fishburn. Probabilistic social choice based on simple voting comparisons. *The Review of Economic Studies*, 51(4):683–692, 1984.

Z. Gao, J. Chang, W. Zhan, O. Oertell, G. Swamy, K. Brantley, T. Joachims, D. Bagnell, J. D. Lee, and W. Sun. Rebel: Reinforcement learning via regressing relative rewards. *Advances in Neural Information Processing Systems*, 37:52354–52400, 2024.

Z. Gao, W. Zhan, J. D. Chang, G. Swamy, K. Brantley, J. D. Lee, and W. Sun. Regressing the relative future: Efficient policy optimization for multi-turn rlhf, 2025. URL https://arxiv.org/abs/2410.04612.

P. Gölz, N. Haghtalab, and K. Yang. Distortion of ai alignment: Does preference optimization optimize for preferences? *arXiv preprint arXiv:2505.23749*, 2025.

A. Gunjal, A. Wang, E. Lau, V. Nath, Y. He, B. Liu, and S. Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

D. Gupta, A. Fisch, C. Dann, and A. Agarwal. Mitigating preference hacking in policy optimization with pessimism. *arXiv preprint arXiv:2503.06810*, 2025.

D. Halpern, E. Micha, A. D. Procaccia, and I. Shapira. Pairwise calibrated rewards for pluralistic alignment. *arXiv preprint arXiv:2506.06298*, 2025.

E. Hazan. Introduction to online convex optimization, 2023. URL https://arxiv.org/abs/1909.05207.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.

B. Heymann. Adaptive preference aggregation. *arXiv preprint arXiv:2503.10215*, 2025.

S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pages 267–274, 2002.

G. Kreweras. Aggregation of preference orderings. In *Mathematics and Social Sciences I: Proceedings of the seminars of Menthon-Saint-Bernard, France (1–27 July 1960) and of Gösing, Austria (3–27 July 1962)*, pages 73–79, 1965.

T. Li, W.-L. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez, and I. Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 34209–34231. PMLR, 2025. URL https://proceedings.mlr.press/v267/li25h.html.

S. Lin, J. Hilton, and O. Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2021.

K. Liu, Q. Long, Z. Shi, W. J. Su, and J. Xiao. Statistical impossibility and possibility of aligning llms with human preferences: From condorcet paradox to nash equilibrium. *arXiv preprint arXiv:2503.10990*, 2025.

R.-R. Maura-Rivero, M. Lanctot, F. Visin, and K. Larson. Jackpot! alignment as a maximal lottery. *arXiv preprint arXiv:2501.19266*, 2025.

R. Munos, M. Valko, D. Calandriello, M. G. Azar, M. Rowland, Z. D. Guo, Y. Tang, M. Geist, T. Mesnard, A. Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 18, 2023.

Y. Nesterov. *Lectures on Convex Optimization*. Springer Optimization and Its Applications, 2 edition, 2018. doi: 10.1007/978-3-319-91578-4.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

A. D. Procaccia and J. S. Rosenschein. The distortion of cardinal preferences in voting. In *International Workshop on Cooperative Information Agents*, pages 317–331. Springer, 2006.

Qwen, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

A. Rame, G. Couairon, C. Dancette, J.-B. Gaya, M. Shukor, L. Soulier, and M. Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36:71095–71134, 2023.

S. S. Ramesh, Y. Hu, I. Chaimalas, V. Mehta, P. G. Sessa, H. Bou Ammar, and I. Bogunovic. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems*, 37:37100–37137, 2024.

C. Rosset, C.-A. Cheng, A. Mitra, M. Santacroce, A. Awadallah, and T. Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

A. Roth. Learning in games (and games in learning). https://www.cis.upenn.edu/~aaroth/GamesInLearning.pdf, 2023. Lecture notes.

A. Shirali, A. Nasr-Esfahany, A. Alomar, P. Mirtaheri, R. Abebe, and A. Procaccia. Direct alignment with heterogeneous preferences. *arXiv preprint arXiv:2502.16320*, 2025.

A. Siththaranjan, C. Laidlaw, and D. Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*, 2023.

Y. Song, G. Swamy, A. Singh, J. Bagnell, and W. Sun. The importance of online data: Understanding preference fine-tuning via coverage. *Advances in Neural Information Processing Systems*, 37: 12243–12270, 2024.

T. Sorensen, J. Moore, J. Fisher, M. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

L. Sutawika, G. Swamy, Z. S. Wu, and G. Neubig. Gained in translation: Privileged pairwise judges enhance multilingual reasoning, 2026. URL https://arxiv.org/abs/2601.18722.

G. Swamy, C. Dann, R. Kidambi, Z. S. Wu, and A. Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

G. Swamy, S. Choudhury, W. Sun, Z. S. Wu, and J. A. Bagnell. All roads lead to likelihood: The value of reinforcement learning in fine-tuning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sCL5mSTpKm.

A. Tversky. Intransitivity of preferences. *Psychological review*, 76(1):31, 1969.

V. Viswanathan, Y. Sun, X. Kong, M. Cao, G. Neubig, and S. Wu. Checklists are better than reward models for aligning language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3824. URL https://papers.nips.cc/paper_files/paper/2025/hash/a6837c1dd021f76f1b4098e3722052a8-Abstract-Conference.html.

J. von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928. doi: 10.1007/BF01448847.

H. Wang, Y. Lin, W. Xiong, R. Yang, S. Diao, S. Qiu, H. Zhao, and T. Zhang. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571*, 2024a.

H. Wang, W. Xiong, T. Xie, H. Zhao, and T. Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*, 2024b.

K. Wang, R. Kidambi, R. Sullivan, A. Agarwal, C. Dann, A. Michi, M. Gelmi, Y. Li, R. Gupta, A. Dubey, et al. Conditional language policy: A general framework for steerable multi-objective finetuning. *arXiv preprint arXiv:2407.15762*, 2024c.

F. Wu, X. Huang, W. Xuan, Z. Zhang, Y. Xiao, G. Wan, X. Li, B. Hu, P. Xia, J. Leskovec, et al. Multiplayer nash preference optimization. *arXiv preprint arXiv:2509.23102*, 2025.

J. Xiao, Z. Shi, K. Liu, Q. Long, and W. J. Su. Theoretical tensions in rlhf: Reconciling empirical success with inconsistencies in social choice theory. *arXiv preprint arXiv:2506.12350*, 2025.

N. Xiong and A. Singh. Projection optimization: A general framework for multi-objective and multi-group rlhf. *arXiv preprint arXiv:2502.15145*, 2025.

Y. Xu, L. Ruis, T. Rocktäschel, and R. Kirk. Investigating non-transitivity in llm-as-a-judge, 2025. URL https://arxiv.org/abs/2502.14074.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, and Z. Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Y. Yue, J. Broder, R. Kleinberg, and T. Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

W. Zhao, X. Ren, J. Hessel, C. Cardie, Y. Choi, and Y. Deng. WildChat: 1m ChatGPT interaction logs in the wild. In *International Conference on Learning Representations*, 2024.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Y. Zhong, C. Ma, X. Zhang, Z. Yang, H. Chen, Q. Zhang, S. Qi, and Y. Yang. Panacea: Pareto alignment via preference adaptation for llms. *Advances in Neural Information Processing Systems*, 37:75522–75558, 2024.

J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

J. P. Zhou, K. Z. Luo, J. Gu, J. Yuan, K. Q. Weinberger, and W. Sun. Orchestrating llms with different personalizations. *arXiv preprint arXiv:2407.04181*, 2024a.

Z. Zhou, J. Liu, J. Shao, X. Yue, C. Yang, W. Ouyang, and Y. Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10586–10613, 2024b.

B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pages 1433–1438, 2008.

A Related Work

Multi-Objective Preference Fine-Tuning. A variety of approaches to multi-objective PFT have been explored in the literature, including minimax game formulations [Chakraborty et al., 2024], distance-based optimization [Wang et al., 2024a, Xiong and Singh, 2025], policy aggregation [Zhou et al., 2024a, A Alamdari et al., 2024, Heymann, 2025], distributional reward modeling [Siththaranjan et al., 2023, Halpern et al., 2025], explicit multi-objective reward modeling [Rame et al., 2023, Wang et al., 2024b], multi-objective reinforcement learning [Zhong et al., 2024, Wang et al., 2024c], and multi-objective direct preference optimization [Chidambaram et al., 2025, Zhou et al., 2024b, Shirali et al., 2025, Ramesh et al., 2024]. However, all of these approaches implicitly assume that along each objective, preferences are transitive and there is a well-defined optimal policy, making it difficult to directly apply them to the more general setting we study in our work.

Intransitive Preferences in Preference Fine-Tuning. Drawing on ideas from social choice theory, several authors have pointed out that care is required when learning from intransitive, *aggregate* preferences to avoid distorting the preferences of the underlying population of raters [Conitzer et al., 2024, Gölz et al., 2025, Maura-Rivero et al., 2025, Liu et al., 2025, Xiao et al., 2025]. However, even when preference feedback is sourced from an LLM rather than a population of raters, intransitivity still frequently arises [Xu et al., 2025], and can therefore stymie policy training. To deal with the fact that there is no well-defined optimal policy under intransitive preferences, a variety of authors have proposed to reframe PFT as a zero-sum game, in which one attempts to find the policy that does the best against its worst-case comparator [Munos et al., 2023, Swamy et al., 2024, Rosset et al., 2024, Sutawika et al., 2026]. These ideas build on a rich history of similar techniques in the social choice theory [Kreweras, 1965, Fishburn, 1984] and bandit [Yue et al., 2012, Dudík et al., 2015] literature. Wu et al. [2025] extend the two-player Nash-equilibrium formulation to the multi-player setting. Azar et al. [2024] also considered scalar general preferences but their motivation is instability and sample inefficiency. While promising, all of these approaches assume the learner will be evaluated along a single objective. In contrast, we focus on the multi-objective setting, requiring nontrivial algorithmic innovation.

More closely to our work, several authors have explored handling multiple, potentially intransitive preferences. Bhatia et al. [2020] pioneered the theoretical study of multi-dimensional intransitive preferences by proposing the Blackwell Winner, but stopped short of deriving a scalable algorithm capable of training modern LLMs as we do. One of the key technical innovations of the MaxEntBW is the use of entropy regularization to avoid the adversarial training Bhatia et al. [2020]'s solution concept implicitly requires. Agnihotri et al. [2025] take a constrained optimization perspective on multi-objective PFT, but essentially sidestep intransitivity by only optimizing the preference against a fixed reference policy, reducing the problem to standard multi-objective RL. In contrast, we directly confront the challenges of multiple intransitive preferences. Perhaps most similarly to our work, Gupta et al. [2025] consider optimizing against multiple judges to deal with poor out-of-distribution judge generalization, deriving a scalable algorithm via the use of a variational relaxation. We implement a similar algorithm as one of our baselines and find that it performs worse than our proposed algorithm, `PROSPER`.

RL from AI Feedback. Our work falls within a growing body of research that leverages AI-generated feedback to guide reinforcement learning. In particular, we elicit AI feedback from judge models that are fed collections of rubrics [Cui et al., 2023, Gunjal et al., 2025, Viswanathan et al., 2025], similar to the "constitutional AI" paradigm used by frontier labs [Bai et al., 2022]. Among rubric-based methods, Cui et al. [2023] evaluate responses using four global criteria rather than prompt-specific checklists, an approach shown to be less effective [Viswanathan et al., 2025], and Gunjal et al. [2025] focus on the RLVR setting, whereas our work targets general chat and instruction following. Consequently, we primarily compare against Viswanathan et al. [2025] in our work. While Viswanathan et al. [2025] employ prompt-specific checklists, they scalarize these per-item rewards into a single score with LLM-generated weights, which can lead to intransitivity. In contrast, we dynamically reweight checklist items based on where the current model is weakest and use game-theoretic techniques to more robustly handle per-item intransitivity. We find that these differences lead to better model performance on both instruction following and general chat benchmarks.

B Supplementary Results

We provide the benchmark performance on Qwen2.5-3B-Instruct in Table 3 and Table 4.

Model	Arena-Hard		AlpacaEval	
	Vanilla	Style-Controlled	Vanilla	Length-Controlled
Qwen2.5-3B-Instruct	20.6	19.6	24.1	12.79
+ RLCF	22.8	22.9	26.2	11.62
+ PROSPER-JC	22.9	23.4	24.7	7.4
+ PROSPER-VB	<u>25.1</u>	<u>23.6</u>	<u>34.8</u>	<u>19.99</u>
+ PROSPER	**26.0**	**24.6**	**35.8**	**20.94**

Table 3: **Model performances on instruction-following and general-chat benchmarks (fine tuning Qwen2.5-3B-Instruct).** Highest score presented in **bold** and second highest <u>underlined</u>. Across instruction-following and preference-alignment benchmarks (AlpacaEval 2.0 and Arena-Hard) evaluated using gpt-5-mini as the judge, PROSPER consistently outperforms the PROSPER ablations (joint-check and variational-bound), while also improving over the RLCF baseline.

Model	IFEval	MMLU	ARC	HellaSwag	TruthfulQA	Avg
Qwen2.5-3B-Instruct	**59.70**	**56.45**	<u>79.43</u>	**67.68**	**64.62**	**65.58**
+ RLCF	58.04	56.28	79.35	66.88	62.17	64.54
+ PROSPER-JC	53.24	56.16	79.27	66.87	61.69	63.45
+ PROSPER-VB	57.30	<u>56.34</u>	**79.52**	<u>67.36</u>	<u>62.67</u>	64.64
+ PROSPER	<u>58.78</u>	56.31	79.27	67.14	62.18	<u>64.74</u>

Table 4: **Model performance on instruction, knowledge, and commonsense benchmarks.** The highest score is shown in **bold** and the second-highest is <u>underlined</u>. Across these benchmarks, post-training methods incur small performance drops; among them, PROSPER ranks first on Qwen2.5-3B-Instruct (by average score).

Also, we provide the results on the checkpoint released by Viswanathan et al. [2025].

Model	Arena-Hard		AlpacaEval	
	Vanilla	Style-Controlled	Vanilla	Length-Controlled
Qwen2.5-7B-Instruct	42.4	44.2	37.1	25.32
+ RLCF (Released)	42.2	44.6	42.3	28.08

Model	IFEval	MMLU	ARC	HellaSwag	TruthfulQA	Avg
Qwen2.5-7B-Instruct	71.16	63.39	85.23	74.54	68.42	72.55
+ RLCF (Released)	67.83	63.16	85.41	74.51	67.44	71.67

C Detailed Proof in Section 3.1

To prove Theorem 3.2, we break the proof into three different parts:

1. Concavity of the optimization objective.
2. Subgradient computation and estimation.
3. Convergence result.

C.1 Concavity of the Optimization Objective

Recall that the optimization objective is

$$\max_{\pi\in\Pi}\ \min_{w:\mathcal{X}\to\Delta_{m(x)}}\ \min_{\pi'\in\Pi}\ \mathbb{E}_x\big[\langle w(x), \mathcal{P}(\pi \succ \pi' \mid x)\rangle + \ \beta\,\mathbb{D}_{\mathrm{KL}}\big(\pi'(x)\,\|\,\pi_{\mathsf{ref}}(x)\big)\big].$$

Using the results from Ziebart et al. [2008], π' has a closed-form solution, and we can then simplify the objective to

$$\max_{\pi\in\Pi}\ \min_{w:\mathcal{X}\to\Delta_{m(x)}}\ \mathbb{E}_x[-\beta\log Z(w,\pi|x)],$$

where

$$Z(w,\pi\mid x) = \mathbb{E}_{y'\sim\pi_{\mathsf{ref}}(x)}[\exp(-\langle w(x), \mathcal{P}_\pi(y'\mid x)\rangle/\beta)].$$

To further simplify the objective, we invoke the following lemma:

Lemma C.1. *$-\beta\log Z(w,\pi\mid x)$ is concave in $w(x)$ and π.*

Proof. By Holder's inequality, for any w_1, w_2 and $\theta\in[0,1]$,

$$\begin{aligned}
Z(\theta w_1 + (1-\theta)w_2, \pi\mid x) &= \mathbb{E}_{y'\sim\pi_{\mathsf{ref}}}[\exp(-\langle w_1(x), \mathcal{P}_\pi(y'\mid x)\rangle)^\theta\cdot\exp(-\langle w_2(x), \mathcal{P}_\pi(y'\mid x)\rangle)^{1-\theta}]\\
&\le \mathbb{E}_{y'\sim\pi_{\mathsf{ref}}}[\exp(-\langle w_1(x), \mathcal{P}_\pi(y'\mid x)\rangle)]^\theta\cdot\mathbb{E}_{y'\sim\pi_{\mathsf{ref}}}[\exp(-\langle w_2(x), \mathcal{P}_\pi(y'\mid x)\rangle)]^{1-\theta}\\
&= Z(w_1,\pi\mid x)^\theta Z(w_2,\pi\mid x)^{1-\theta}.
\end{aligned}$$

Taking $-\beta\log(\cdot)$ on both sides, we have

$$-\beta\log Z(\theta w_1 + (1-\theta)w_2,\pi\mid x) \ge \theta(-\beta\log Z(w_1,\pi\mid x)) + (1-\theta)(-\beta\log Z(w_2,\pi\mid x)).$$

Thus, $-\beta\log Z(w,\pi\mid x)$ is concave in $w(x)$. A similar argument applies for π. □

By using a concave-form statement of Bauer's maximum principle Bauer [1958] as shown in theorem C.2, it suffices to enumerate only the unit vector in $\Delta_{m(x)}$ for each w.

Theorem C.2 (Bauer's maximum principle, concave form)**.** *Let C be a nonempty compact convex subset of a (Hausdorff) locally convex topological vector space and let $\mathcal{C}$ be the set of extreme points in C. Let $f: C\to\mathbb{R}$ be concave and lower semicontinuous (in particular, continuous suffices). Then f attains its minimum at an extreme point of C:*

$$\min_{x\in C} f(x) \ = \ \min_{x\in\mathcal{C}} f(x).$$

By Theorem C.2, the optimization problem becomes $\max_{\pi\in\Pi}\mathbb{E}_x[-\beta\log Z^{k^\star(x)}(\pi|x)]$. Now we show

$$\mathbb{E}_x\Big[-\beta\log Z^{k^\star(x)}(\pi|x)\Big] = \mathbb{E}_x\Big[\min_{k\in[m(x)]} -\beta\log Z^k(\pi|x)\Big]$$

is concave in π. It suffices to show $\min_{k\in[m(x)]}-\beta\log Z^k(\pi|x)$ is concave in π. This is because $-\beta\log Z^k$ is concave in π and the minimum of $m(x)$ concave functions is a concave function. Then the question boils down to **(sub)gradient computation and estimation**.

C.2 Subgradient computation and estimation

By a classical result in convex analysis Nesterov [2018], we have:

Lemma C.3. *Given any integer N, Let $f(x) = \max_{i\in[N]} f_i(x)$, where each f_i is convex. Then the subdifferential of f at x satisfies*

$$\partial f(x) = \operatorname{conv}\{\partial f_i(x) : i \in \arg\max_i f_i(x)\}.$$

Lemma C.4.

$$g^k(z \mid \pi, x) \triangleq [\nabla_\pi - \beta \log Z^k(\pi \mid x)](z) = \frac{\mathbb{E}_{y'\sim\pi_{\text{ref}}}[\mathcal{P}^k(z \succ y' \mid x)\exp(-\mathbb{E}_{y\sim\pi}[\mathcal{P}^k(y \succ y' \mid x)]/\beta)]}{\mathbb{E}_{y'\sim\pi_{\text{ref}}}[\exp(-\mathbb{E}_{y\sim\pi}[\mathcal{P}^k(y \succ y' \mid x)]/\beta)]}.$$

Proof. By calculation. □

Lemma C.5. *The gradient of the objective function $-\beta \log Z^{k^\star(x)}(\pi \mid x)$ is*

$$g(z \mid \pi, x) = \frac{\mathbb{E}_{y'\sim\pi_{\text{ref}}}[\mathcal{P}^{k^\star(x)}(z \succ y' \mid x)\exp(-\mathbb{E}_{y\sim\pi}[\mathcal{P}^{k^\star(x)}(y \succ y' \mid x)]/\beta)]}{\mathbb{E}_{y'\sim\pi_{\text{ref}}}[\exp(-\mathbb{E}_{y\sim\pi}[\mathcal{P}^{k^\star(x)}(y \succ y' \mid x)]/\beta)]}.$$

Proof. Combining Lemma C.3 and Lemma C.4 yields the conclusion. □

C.3 Convergence Result

For notational convenience, we write π_t in place of π_{θ_t} and $g_t(x, z)$ in place of $g(z \mid \pi_{\theta_t}, x)$. In the following, we define $f_t(x, z) := \frac{1}{\eta}\log\frac{\pi_{t+1}(z|x)}{\pi_t(z|x)}$.

Lemma C.6. *Consider any $t \in [T]$. Define $\Delta(x, z) = f_t(x, z) - g_t(x, z)$. Define $\Delta_{\pi_t}(x) = \mathbb{E}_{z\sim\pi_t(x)}\Delta(x, z)$ and $\Delta_{\pi_{\text{ref}}}(x) = \mathbb{E}_{z\sim\pi_{\text{ref}}(x)}\Delta(x, z)$. Under Assumption 3.1, for all t, we have the following:*

$$\mathbb{E}_{x,z\sim\pi_t(x)}\left(f_t(x, z) - g_t(x, z) - \Delta_{\pi_t}(x)\right)^2 \le \epsilon, \tag{13}$$

$$\mathbb{E}_{x,z\sim\pi_{\text{ref}}(x)}\left(f_t(x, z) - g_t(x, z) - \Delta_{\pi_{\text{ref}}}(x)\right)^2 \le \epsilon, \tag{14}$$

$$\mathbb{E}_x\left(\Delta_{\pi_t}(x) - \Delta_{\pi_{\text{ref}}}(x)\right)^2 \le \epsilon. \tag{15}$$

Proof. From Assumption 3.1, we have:

$$\begin{aligned}
&\mathbb{E}_{x,z_1\sim\pi_t,z_2\sim\pi_{\text{ref}}}\left(f_t(x, z_1) - \Delta_{\pi_t}(x) - g_t(x, z_1) - (f_t(x, z_2) - \Delta_{\pi_{\text{ref}}}(x) - g_t(x, z_2)) + \Delta_{\pi_t}(x) - \Delta_{\pi_{\text{ref}}}(x)\right)^2 \\
=&\mathbb{E}_{x,z_1\sim\pi_t}\left(f_t(x, z_1) - \Delta_{\pi_t}(x) - g_t(x, z_1)\right)^2 + \mathbb{E}_{x,z_2\sim\pi_{\text{ref}}}\left(f_t(x, z_2) - \Delta_{\pi_{\text{ref}}}(x) - g_t(x, z_2)\right)^2 \\
&\quad - 2\mathbb{E}_{x,z_1\sim\pi_t,z_2\sim\pi_{\text{ref}}}\left(f_t(x, z_1) - \Delta_{\pi_t}(x) - g_t(x, z_1)\right)\left(f_t(x, z_2) - \Delta_{\pi_{\text{ref}}}(x) - g_t(x, z_2)\right) \\
&\quad + 2\mathbb{E}_x\left(f_t(x, z_1) - \Delta_{\pi_t}(x) - g_t(x, z_1)\right)\left(\Delta_{\pi_t}(x) - \Delta_{\pi_{\text{ref}}}(x)\right) \\
&\quad - 2\mathbb{E}_{x,z_2\sim\pi_{\text{ref}}}\left(\Delta_{\pi_t}(x) - \Delta_{\pi_{\text{ref}}}(x)\right)\left(f_t(x, z_2) - \Delta_{\pi_{\text{ref}}}(x) - g_t(x, z_2)\right) + \mathbb{E}_x\left(\Delta_{\pi_t}(x) - \Delta_{\pi_{\text{ref}}}(x)\right)^2 \\
=&\mathbb{E}_{x,z_1\sim\pi_t}\left(f_t(x, z_1) - \Delta_{\pi_t}(x) - g_t(x, z_1)\right)^2 + \mathbb{E}_{x,z_2\sim\pi_{\text{ref}}}\left(f_t(x, z_2) - \Delta_{\pi_{\text{ref}}}(x) - g_t(x, z_2)\right)^2 \\
&\quad + \mathbb{E}_x\left(\Delta_{\pi_t}(x) - \Delta_{\pi_{\text{ref}}}(x)\right)^2 \le \epsilon.
\end{aligned}$$

□

By definition, we have $\Delta(x, z) = \frac{1}{\eta}\log\frac{\pi_{t+1}(z|x)}{\pi_t(z|x)} - g_t(x, z)$. Taking exp on both sides, we have

$$\pi_{t+1}(z \mid x) = \pi_t(z \mid x)\exp(\eta(g_t(x, z)+\Delta(x, z))) = \frac{\pi_t(z \mid x)\exp(\eta(g_t(x, z) + \Delta(x, z) - \Delta_{\pi_{\text{ref}}}(x)))}{\exp(-\eta\Delta_{\pi_{\text{ref}}}(x))}.$$

Denote $h_t(x, z) := g_t(x, z) + \Delta(x, z) - \Delta_{\pi_{\text{ref}}(x)}$ and the advantage $A_t(x, z) = h_t(x, z) - \mathbb{E}_{z'\sim\pi_t(x)}[h_t(x, z')]$. We can rewrite the update rule as

$$\pi_{t+1}(z \mid x) \propto \pi_t(z \mid x)\exp(\eta A_t(x, z)).$$

Lemma C.7. *Assume* $\max_{x,y,t} |A_t(x,y)| \le A \in \mathbb{R}^+$*, and* $\pi_0(x)$ *is uniform over* $\mathcal{Y}$*. Then with*

$$\eta = \sqrt{\ln(|\mathcal{Y}|)/(A^2 T)},$$

for the sequence of policies computed by PROSPER*, we have:*

$$\forall \pi, x : \sum_{t=0}^{T-1} \mathbb{E}_{y\sim\pi_t(x)} A_t(x,y) \le 2A\sqrt{\ln(|\mathcal{Y}|)T}.$$

Proof. Start with

$$\pi_{t+1}(y \mid x) = \pi_t(y \mid x) \exp(\eta A_t(x,y))/Z_t(x),$$

where $Z_t(x)$ is the normalization constant. Taking log on both sides and add $\mathbb{E}_{y\sim\pi_t(x)}$, we have:

$$-\mathbb{D}_{\mathrm{KL}}(\pi(x)\|\pi_{t+1}(x)) = -\mathbb{D}_{\mathrm{KL}}(\pi(x)\|\pi_t(x)) + \eta \mathbb{E}_{y\sim\pi(x)} A_t(x,y) - \mathbb{E}_{y\sim\pi(x)} \ln Z_t(x).$$

Rearranging terms, we get:

$$-\mathbb{D}_{\mathrm{KL}}(\pi(x)\|\pi_t(x)) + \mathbb{D}_{\mathrm{KL}}(\pi(x)\|\pi_{t+1}(x)) = \mathbb{E}_{y\sim\pi(x)}[-\eta A_t(x,y) + \ln Z_t(x)].$$

For $\ln Z_t(x)$, using the condition that $\eta \le 1/A$, we have $\eta A_t(x,y) \le 1$, which allows us to use the inequality $\exp(x) \le 1 + x + x^2$ for any $x \le 1$, which leads to the following inequality:

$$\begin{aligned}
\ln Z_t(x) &= \ln\left(\mathbb{E}_{y\sim\pi_t(x)} \exp(\eta A_t(x,y))\right) \\
&\le \ln\left(\sum_y \pi_t(y \mid x)\left(1 + \eta A_t(x,y) + \eta^2 A_t(x,y)^2\right)\right) \\
&\le \ln\left(1 + 0 + \eta^2 A^2\right) \le \eta^2 A^2,
\end{aligned}$$

where the last inequality uses $\ln(1+x) \le x$, and we used the fact that $\mathbb{E}_{y\sim\pi_t(x)} A_t(x,y) = 0$ due to the definition of advantage A_t.

Thus, we have:

$$-\mathbb{D}_{\mathrm{KL}}(\pi(x)\|\pi_t(x)) + \mathbb{D}_{\mathrm{KL}}(\pi(x)\|\pi_{t+1}(x)) \le -\eta \mathbb{E}_{y\sim\pi(x)}[A_t(x,y)] + \eta^2 A^2.$$

Sum over all iterations and do the telescoping sum, we get:

$$\sum_{t=0}^{T-1} \mathbb{E}_{y\sim\pi(x)} A_t(x,y) \le \mathbb{D}_{\mathrm{KL}}(\pi(x)\|\pi_0(x))/\eta + T\eta A^2 \le \ln(|\mathcal{Y}|)/\eta + T\eta A^2.$$

With $\eta = \sqrt{\ln(|\mathcal{Y}|)/(A^2 T)}$, we conclude the proof. □

Now we can prove the Theorem 3.2 formally:

Theorem 3.2. *Under Assumption 3.1, after* T *iterations, there exists some* $\hat{\pi} \in \{\pi_{\theta_1}, \cdots, \pi_{\theta_T}\}$ *such that* $V(\pi^\star) - V(\hat{\pi}) \le O(\sqrt{1/T} + \sqrt{C_{\pi_{\mathsf{ref}}\to\pi^\star}\epsilon})$*, where* $\pi^\star \in \arg\max_{\pi\in\Pi} V(\pi)$ *denotes an arbitrary MaxEntBW.*

Proof. Fix a comparator policy $\pi^\star$. We begin by considering the average performance gap between $\pi^\star$ and π_t over $t = 0, 1, \ldots, T-1$:

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}(V(\pi^\star) - V(\pi_t)) &\le \frac{1}{T}\sum_{t=0}^{T-1}(\langle \nabla V(\pi_t), \pi^\star - \pi_t\rangle) \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}_{x,y\sim\pi^\star(x)}[g_t(x,y)] - \mathbb{E}_{x,y\sim\pi_t(x)}[g_t(x,y)]\right) \\
&= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{x,y\sim\pi^\star(x)}\left(A^{\pi_t}(x,y)\right),
\end{aligned}$$

where the first inequality holds because of concavity and we define the real advantage $A^{\pi_t}(x,y) := g_t(x,y) - \mathbb{E}_{z\sim\pi_t(x)} g_t(x,z)$. Continuing, we have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{x,y\sim\pi^\star(x)}\left(A^{\pi_t}(x,y)\right) = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{x,y\sim\pi^\star(x)}\left(A_t(x,y)\right) + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{x,y\sim\pi^\star(x)}\left(A^{\pi_t}(x,y) - A_t(x,y)\right)$$

$$\leq 2A\sqrt{\frac{\ln(|\mathcal{Y}|)}{T}} + \frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\mathbb{E}_{x,y\sim\pi^\star(x)}\left(A^{\pi_t}(x,y) - A_t(x,y)\right)^2},$$

where the last inequality uses Lemma C.7. We now just need to bound $\mathbb{E}_{x,y\sim\pi^\star(x)}\left(A^{\pi_t}(x,y) - A_t(x,y)\right)^2$.

$$\mathbb{E}_{x,y\sim\pi^\star(x)}\left(A^{\pi_t}(x,y) - A_t(x,y)\right)^2 = \mathbb{E}_x\mathbb{E}_{y\sim\pi_{\mathsf{ref}}(x)}\left(\frac{\pi^\star(y\mid x)}{\pi_{\mathsf{ref}}(y\mid x)}(A^{\pi_t}(x,y) - A_t(x,y))^2\right)$$

$$\leq C_{\pi_{\mathsf{ref}}\to\pi^\star}\cdot\mathbb{E}_{x,y\sim\pi_{\mathsf{ref}}(x)}\left(A^{\pi_t}(x,y) - A_t(x,y)\right)^2$$

where the last inequality uses the definition of *concentrability coefficient* $C_{\pi_{\mathsf{ref}}\to\pi^\star}$, which is

$$C_{\pi_{\mathsf{ref}}\to\pi} = \max_{x,y}\frac{\pi(y\mid x)}{\pi_{\mathsf{ref}}(y\mid x)}.$$

We now bound $\mathbb{E}_{x,y\sim\pi_{\mathsf{ref}}(x)}\left(A^{\pi_t}(x,y) - A_t(x,y)\right)^2$. Recall the definition of A_t from Lemma C.7:

$$\begin{aligned}\mathbb{E}_{x,y\sim\pi_{\mathsf{ref}}(x)}\left(A^{\pi_t}(x,y) - A_t(x,y)\right)^2 &= \mathbb{E}_{x,y,y'\sim\pi_t(x)}\left(\hat{g}_t(x,y) - \mathbb{E}_{y'\sim\pi_t(x)}\hat{g}_t(x,y') - h_t(x,y) + \mathbb{E}_{y'\sim\pi_t(x)}h_t(x,y')\right)^2\\ &\leq 2\mathbb{E}_{x,y\sim\pi_{\mathsf{ref}}(x)}\left(\hat{g}_t(x,y) - h_t(x,y)\right)^2 + 2\mathbb{E}_{x,y'\sim\pi_t(x)}\left(\mathbb{E}_{y\sim\pi_t(x)}[\hat{g}_t(x,y) - h_t(x,y)]\right)^2\end{aligned}$$

Recall that $h_t(x,y) = g(x,y) + \Delta(x,y) - \Delta_{\pi_{\mathsf{ref}}}(x)$, and from Lemma C.6, we can see that:

$$\mathbb{E}_{x,y\sim\pi_{\mathsf{ref}}(x)}\left(g_t(x,y) - h_t(x,y)\right)^2 = \mathbb{E}_{x,y\sim\pi_{\mathsf{ref}}(x)}\left(\Delta(x,y) - \Delta_{\pi_{\mathsf{ref}}}(x)\right)^2 \leq \epsilon.$$

For

$$\mathbb{E}_x\mathbb{E}_{y'\sim\pi_t(x)}\left(g_t(x,y') - h_t(x,y')\right)^2,$$

we have:

$$\begin{aligned}\mathbb{E}_x\mathbb{E}_{y'\sim\pi_t(x)}\left(\hat{g}_t(x,y') - h_t(x,y')\right)^2 &= \mathbb{E}_x\mathbb{E}_{y'\sim\pi_t(x)}\left(\Delta_{\pi_t}(x,y') - \Delta_{\pi_{\mathsf{ref}}}(x)\right)^2\\ &= \mathbb{E}_x\mathbb{E}_{y'\sim\pi_t(x)}\left(\Delta(x,y') - \Delta_{\pi_t}(x) + \Delta_{\pi_t}(x) - \Delta_{\pi_{\mathsf{ref}}}(x)\right)^2\\ &\leq 2\mathbb{E}_x\mathbb{E}_{y'\sim\pi_t(x)}\left(\Delta(x,y') - \Delta_{\pi_t}(x)\right)^2 + 2\mathbb{E}_x\left(\Delta_{\pi_t}(x) - \Delta_{\pi_{\mathsf{ref}}}(x)\right)^2\\ &\leq 4\epsilon,\end{aligned}$$

where the last inequality uses Lemma C.6 again. This step relies on the fact that one of the samples is always on-policy, i.e., from π_t.

Combine things together, we can conclude that:

$$\mathbb{E}_x\mathbb{E}_{y\sim\pi^\star(x)}\left(A^{\pi_t}(x,y) - A_t(x,y)\right)^2 \leq C_{\pi_{\mathsf{ref}}\to\pi^\star}(10\epsilon).$$

We can conclude:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\left(V(\pi^\star) - V(\pi_t)\right) &\leq 2A\sqrt{\frac{\ln|\mathcal{Y}|}{T}} + \frac{1}{T}\sum_t\sqrt{C_{\pi_{\mathsf{ref}}\to\pi^\star}10\epsilon}\\ &= 2A\sqrt{\frac{\ln|\mathcal{Y}|}{T}} + \sqrt{C_{\pi_{\mathsf{ref}}\to\pi^\star}10\epsilon}.\end{aligned}$$

□

D Implementation Details

D.1 Implementation Details and Hyperparameters

Our algorithm follows an off-policy training paradigm, where each epoch is partitioned into three distinct phases: response generation, preference scoring, and policy optimization. We execute two training epochs per model. The following sections detail the hyperparameter configurations and computational requirements for each phase.

Response Generation. In each epoch, we sample $K = 8$ responses per prompt from the current policy. The sampling parameters and the computational resources utilized for this stage are summarized in Table 5.

Table 5: Hyperparameters and compute configuration for response generation.

Parameter	**Value**
Temperature	0.8
Top-p	0.9
Max new tokens	2048
Hardware	$1 \times (16 \times$ NVIDIA H100 80GB$)$
Total compute time	≈ 2 hours per model per epoch

Score Generation. We utilize an LLM-as-a-judge framework to evaluate the generated response pairs. Each pair is integrated into a structured evaluation template, and any resulting sequences exceeding a length of 4096 tokens are excluded to maintain consistency. To ensure the robustness of the preference signals, we sample 10 independent judgments per (prompt, response pair). To mitigate inherent positional bias, we implement a symmetric evaluation protocol: the presentation order of the responses is permuted for half of the samples (5 out of 10), and the corresponding scores are reversed before computing the final average. Detailed scoring parameters are summarized in Table 6.

Table 6: Hyperparameters and compute configuration for score generation.

Parameter	**Value**
Temperature	0.8
Top-p	0.9
Top-k	20
Max new tokens	256
Selection pairs	4
Base pairs	2
Current pairs	2
Hardware	$1 \times (16 \times$ NVIDIA H100 80GB$)$
Total compute time	≈ 50 hours per model per epoch

Training and Optimization. Following the derivation of game-theoretic training signals from the preference scores, we filter the dataset to retain the most informative signals. Specifically, we expand the dataset by constructing all possible pairwise comparisons within each prompt and then apply a global filter based on the score gap. We use a filtration ratio of 0.15 for the first epoch and 0.17 for the second, selected based on performance on the ARENA-HARD benchmark. The policy is then optimized using the hyperparameters listed in Table 7.

Table 7: Hyperparameters and compute configuration for policy training.

Parameter	**Value**
Batch size	128
Max input length	1024
Max sequence length	2048
Optimizer	AdamW
AdamW ϵ	1×10^{-8}
Learning rate	3×10^{-7}
Weight decay	1×10^{-6}
Warmup ratio	0.1
Max gradient norm	1.0
Seed	555134
Hardware	$1 \times (8 \times$ NVIDIA H100 80GB$)$
Total compute time	≈ 3 hours per model per epoch

D.2 Prompts for LLM Judge

You are an expert in comparing the quality of responses and giving preference scores. Based on the provided input instruction and two responses from a worker, assess two responses based on the following criteria:
1. Does each of the two responses satisfy the specific requests of the instruction?
2. Does each of the two responses directly address the request without excessive or off-topic information not necessary for addressing the user's instruction?
3. Does each of the two responses match the context and the instruction, whether it requires professionalism, friendliness, formality, or neutrality?
Accordingly, assign a score (in {{0,1,2,3,4}}) to indicate the preference for the first response over the second assessing how well the first response addresses the instruction compared to the second response.
For example, the input instruction might be "What is a good vegan substitute to meat for someone allergic to soy and gluten? Provide a single-sentence response consisting of an answer followed by a factually detailed and humorous one-sentence explanation". Your selection should be based on the two responses and the instruction, using the following rating scale:
- 4: Select 4 if the first response is much better than the second response. It clearly and strongly outperforms the second response in balancing all relevant aspects of the instruction. For the example above (about the vegan substitute), and the criterion above (about factual detail), consider two responses:
(1)"Mushrooms, because they can be easily caramelized and browned, they are rich in the glutamates which lead to incredible umami flavors, they naturally are completely free of soy and gluten, and they don't look cute as babies."
(2)"Mushrooms are tasty."
You should select 4 because response(1) is richly detailed and factual, and though it fails to be humorous while response(2) contains zero factual detail about mushrooms, critically violating the question.
- 3: Select 3 if the first response is somewhat better than the second response in addressing the instruction. Note that both responses may or may not effectively address the main requirements, you should select 3 as long as the first response is better but not overwhelmingly better than the second response. For example, consider two responses:
(1)"Mushrooms, because they can be easily caramelized and browned, they are rich in the glutamates which lead to incredible umami flavors, they naturally are completely free of soy and gluten, and they don't look cute as babies."
(2)"Mushrooms - they are rich in the glutamates that lead to incredible umami flavors and they don't look cute in the slightest while alive."
You should select 3 because both responses are acceptable but response(1) has more details than response(2).
Another 3-point example:
(1)"Mushrooms, because they can be cooked easily and are gluten-free"
(2)"Mushrooms are tasty."
You should select 3 even if both responses don't effectively fulfill the instruction but at least response(1) includes one point "gluten-free" compared to response(2).
- 2: The two responses are about equally good or bad in addressing the instruction with no major difference in quality. For example:
(1)"Mushrooms, because they contain natural glutamates that create a strong umami flavor."
(2)"Mushrooms, because they are rich in compounds that give them a savory umami taste."
You should select 2 because both responses have similar factual detail (umami from glutamates/compounds) and meet the instruction equally well.
Another 2-point example:
(1)"Mushrooms are tasty."
(2)"Mushrooms, because they can be cooked easily."
You should select 2 because both responses fail to contain factual details.
- 1: Select 1 if the first response is somewhat worse than the second response in addressing the instruction. Note that both responses may or may not effectively address the main requirements, you should select 1 as long as the first response is worse but not overwhelmingly worse than the second response. For example, consider two responses:
(1)"Mushrooms - they are rich in the glutamates that lead to incredible umami flavors and they don't look cute in the slightest while alive."
(2)"Mushrooms, because they can be easily caramelized and browned, they are rich in the glutamates which lead to incredible umami flavors, they naturally are completely free of soy and gluten, and they don't look cute as babies."
You should select 1 because both responses are acceptable but response(1) has less details than response(2).
Another 1-point example:
(1)"Mushrooms are tasty."
(2)"Mushrooms, because they can be cooked easily and are gluten-free"
You should select 1 even if both responses don't effectively fulfill the instruction but response(1) has no factual details at all and response(2)at least includes one point "gluten-free".
- 0: Select 0 if the first response is much worse than the second response. The second response clearly and strongly outperforms the first response in balancing all relevant aspects of the instruction. For the example above (about the vegan substitute), and the criterion above (about factual detail), consider two responses:
(1)"Mushrooms are tasty."
(2)"Mushrooms, because they can be easily caramelized and browned, they are rich in the glutamates which lead to incredible umami flavors, they naturally are completely free of soy and gluten, and they don't look cute as babies."
You should select 0 because response(1) contains zero factual detail about mushrooms, critically violating the question while response(2) is richly detailed and factual, though it fails to be humorous.
Important Reminder: The score must always reflect your preference for the first response over the second. This is a relative judgment — do not evaluate each response in isolation. Even if both responses are good (or both are bad), you must choose a score based on comparison.
Your score can be any number in {{0,1,2,3,4}}. If you are totally confused, return -1 as a default. You should use your judgment to determine the most appropriate score. Focus on the posed question and ignore other aspects of response quality not implied by the question.
First provide a short 1-2 explanation of your choice and why you made it. Then provide your verdict as a single number in {{0,1,2,3,4}}. Use the format:
`{{"explanation":"...", "verdict":"..."}}`

Input:
{prompt}

Generated Text A:
{response_a}

Generated Text B:
{response_b}

Question:
{check}

Preference:

Figure 4: Prompt for generating preference score according to specific criteria.

You are a expert in comparing the quality of responses and giving preference scores. Based on the provided input instruction, the questions below, and two responses from a worker, assess two responses based on the
following criteria:
1. Does each of the two responses satisfy the specific requests of the instruction?
2. Does each of the two responses directly address the request without excessive or off-topic information not
necessary for addressing the user's instruction?
3. Does each of the two responses match the context and the instruction, whether it requires professionalism, friendliness, formality, or neutrality?
Accordingly, assign a score (in {{0,1,2,3,4}}) to indicate the preference for the first response over the second assessing how well the first response addresses the instruction and the questions compared to the second response.
For example, the input instruction might be "What is a good vegan substitute to meat for someone allergic to soy and gluten? Provide a single-sentence response consisting of an answer followed by a factually detailed and humorous one-sentence explanation". Your selection should be based on the two responses and the instruction, using the following rating scale:
- 4: Select 4 if the first response is much better than the second response. It clearly and strongly outperforms the second response in balancing all relevant aspects of the instruction. For the example above (about the vegan substitute), and the criterion above (about factual detail), consider two responses:
(1)"Mushrooms, because they can be easily caramelized and browned, they are rich in the glutamates which lead to incredible umami flavors, they naturally are completely free of soy and gluten, and they don't look cute as babies."
(2)"Mushrooms are tasty."
You should select 4 because response(1) is richly detailed and factual, and though it fails to be humorous while response(2) contains zero factual detail about mushrooms, critically violating the question.
- 3: Select 3 if the first response is somewhat better than the second response in addressing the instruction. Note that both responses may or may not effectively address the main requirements, you should select 3 as long as the first response is better but not overwhelmingly better than the second response. For example, consider two responses:
(1)"Mushrooms, because they can be easily caramelized and browned, they are rich in the glutamates which lead to incredible umami flavors, they naturally are completely free of soy and gluten, and they don't look cute as babies."
(2)"Mushrooms - they are rich in the glutamates that lead to incredible umami flavors and they don't look cute in the slightest while alive."
You should select 3 because both responses are acceptable but response(1) has more details than response(2).
Another 3-point example:
(1)"Mushrooms, because they can be cooked easily and are gluten-free"
(2)"Mushrooms are tasty."
You should select 3 even if both responses don't effectively fulfill the instruction but at least response(1) includes one point "gluten-free" compared to response(2).
- 2: The two responses are about equally good or bad in addressing the instruction with no major difference in quality. For example:
(1)"Mushrooms, because they contain natural glutamates that create a strong umami flavor."
(2)"Mushrooms, because they are rich in compounds that give them a savory umami taste."
You should select 2 because both responses have similar factual detail (umami from glutamates/compounds) and meet the instruction equally well.
Another 2-point example:
(1)"Mushrooms are tasty."
(2)"Mushrooms, because they can be cooked easily."
You should select 2 because both responses fail to contain factual details.
- 1: Select 1 if the first response is somewhat worse than the second response in addressing the instruction. Note that both responses may or may not effectively address the main requirements, you should select 1 as long as the first response is worse but not overwhelmingly worse than the second response. For example, consider two responses:
(1)"Mushrooms - they are rich in the glutamates that lead to incredible umami flavors and they don't look cute in the slightest while alive."
(2)"Mushrooms, because they can be easily caramelized and browned, they are rich in the glutamates which lead to incredible umami flavors, they naturally are completely free of soy and gluten, and they don't look cute as babies."
You should select 1 because both responses are acceptable but response(1) has less details than response(2).
Another 1-point example:
(1)"Mushrooms are tasty."
(2)"Mushrooms, because they can be cooked easily and are gluten-free"
You should select 1 even if both responses don't effectively fulfill the instruction but response(1) has no factual details at all and response(2)at least includes one point "gluten-free".
- 0: Select 0 if the first response is much worse than the second response. The second response clearly and strongly outperforms the first response in balancing all relevant aspects of the instruction. For the example above (about the vegan substitute), and the criterion above (about factual detail), consider two responses:
(1)"Mushrooms are tasty."
(2)"Mushrooms, because they can be easily caramelized and browned, they are rich in the glutamates which lead to incredible umami flavors, they naturally are completely free of soy and gluten, and they don't look cute as babies."
You should select 0 because response(1) contains zero factual detail about mushrooms, critically violating the question while response(2) is richly detailed and factual, and though it fails to be humorous.
Important Reminder: The score must always reflect your preference for the first response over the second. This is a relative judgment — do not evaluate each response in isolation. Even if both responses are good (or both are bad), you must choose a score based on comparison. Consider ALL listed questions collectively (and their importance if specified) when making your decision.
Your score can be any number in {{0,1,2,3,4}}. If you are totally confused, return -1 as a default. You should use your judgment to determine the most appropriate score. Focus on the posed instruction and the full set of questions, and ignore other aspects of response quality not implied by them.
First provide a short 1-2 explanation of your choice and why you made it. Then provide your verdict as a single number in {{0,1,2,3,4}}. Use the format:
`{{"explanation":"...", "verdict":"..."}}`

Input:
{prompt}

Generated Text A:
{response_a}

Generated Text B:
{response_b}

Questions (consider all together):
{check}

Preference:

Figure 5: Prompt for generating preference score according to multiple criterion.

We provide the prompts used for requirement checking in Figure 4 and Figure 5. In Figure 4, each check for a given prompt is provided separately, whereas in Figure 5, all checks for the prompt are provided together. The templates are adapted from those in Viswanathan et al. [2025], modified for pairwise comparison and augmented with explanation prompting to improve scoring quality.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 3.1 state the proposed MaxEntBW solution concept and PROSPER algorithm, and the experimental/theoretical sections substantiate those claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discuss the limitations in the conclusion section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The main theoretical result is stated under Assumption 3.1 and the concentrability condition, with supporting definitions of the objective and gradients. Appendix C gives the concavity lemmas, subgradient calculation, and convergence proof for Theorem 3.2.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper identifies the training data, policy and judge models, benchmark suite, baselines, ablations, sampling setup, optimizer, prompts, and compute configuration in Section 4 and Appendix D. These details are sufficient to reproduce the high-level experimental pipeline.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code and models will be released publicly.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 specifies dataset, model, judge, benchmark, baseline, and ablation choices; Appendix D.1 gives sampling, scoring, filtering, optimizer, sequence-length, seed, and hardware settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Owing to limited computational resources, we did not perform multiple experimental runs. This approach aligns with established literature Feng et al. [2024], given the low variance observed across different seeds.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix D reports hardware and approximate runtime for response generation, score generation, and policy training in Tables 5, 6, and 7, including H100 counts and per-model/per-epoch time. This provides the key compute information needed for reproduction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses existing datasets, public/pretrained models, and AI-generated preference feedback, and it does not report new human-subject or crowdsourcing studies. To the best of the authors' knowledge, the work conforms to the NeurIPS Code of Ethics; anonymity should be maintained in any released supplementary material.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The potential societal impacts are discussed in the Broader Impacts section of the paper.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: As the model is not yet publicly available, the specific safeguards will be detailed and implemented concurrent with its release.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets and models utilized in this study are appropriately cited and credited within the manuscript.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We'll release the trained models upon paper acceptance.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not conduct new crowdsourcing experiments or human-subject studies. It relies on existing datasets and LLM-as-a-judge evaluations; judge prompt templates are provided in Appendix D.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve new human-subject or crowdsourcing experiments, so IRB approval or equivalent human-subject review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are central to the methodology and experiments: the paper fine-tunes Qwen2.5-Instruct models, uses Qwen3-14B as an LLM judge, and includes LLM judge prompts in Appendix D. Section 5 also describes additional LLM use for rubric generation and benchmark judging.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.