

# SEMANTIC-AWARE WASSERSTEIN POLICY REGULARIZATION FOR LARGE LANGUAGE MODEL ALIGNMENT

Byeonghu Na<sup>1</sup>, Hyungho Na<sup>2</sup>, Yeongmin Kim<sup>1</sup>, Suhyeon Jo<sup>1</sup>, HeeSun Bae<sup>1</sup>, Mina Kang<sup>1</sup>, Il-Chul Moon<sup>1,3</sup>

<sup>1</sup>KAIST, <sup>2</sup>UNIST, <sup>3</sup>summary.ai

byeonghu.na@kaist.ac.kr, h.na@unist.ac.kr

{alsdudrla10, suhyeonjo, cat2507, kasongl3, icmoon}@kaist.ac.kr

## ABSTRACT

Large language models (LLMs) are commonly aligned with human preferences using reinforcement learning from human feedback (RLHF). In this method, LLM policies are generally optimized through reward maximization with Kullback-Leibler (KL) divergence regularization of the reference policy. However, KL and its  $f$ -divergence variants only compare token probabilities at identical indices, failing to capture semantic similarity. We propose Wasserstein Policy Regularization (WPR), a semantic-aware regularization for the RLHF framework based on the entropy-regularized Wasserstein distance, which incorporates the geometry of the token space. The dual formulation of the distance expresses the regularization as penalty terms applied to the reward via optimal dual variables, which yield a tractable objective compatible with standard RL algorithms. Empirically, our method outperforms KL- and  $f$ -divergence-based baselines, demonstrating the benefits of semantic-aware policy distances for alignment. Our code is available at <https://github.com/aailab-kaist/WPR>.

## 1 INTRODUCTION

Large language models (LLMs) have achieved remarkable progress in recent years, powering applications ranging from conversational agents to code generation (Touvron et al., 2023; Achiam et al., 2023; Hui et al., 2024). A central challenge in their deployment is aligning model behavior with human preferences. Reinforcement learning from human feedback (RLHF) has emerged as the dominant paradigm for alignment, where models are optimized to better reflect user intent (Christiano et al., 2017; Bai et al., 2022; Ouyang et al., 2022). The standard RLHF pipeline trains a reward model from human preference data and optimizes the LLM policy to maximize reward while remaining close to a supervised fine-tuned reference model (Ouyang et al., 2022). Recent advances such as Direct Preference Optimization (DPO) (Rafailov et al., 2023) and its variants (Azar et al., 2024; Ethayarajh et al., 2024) follow a similar principle, reducing the preference learning to implicit reward maximization with reverse Kullback–Leibler (KL) regularization to maintain the reference policy.

The policy regularization by the KL divergence is widely adopted because the KL divergence can be computed directly from the token probabilities of the reference and the trained models, which is implemented as a penalty on the reward. While KL-based regularization is effective in practice, it exhibits known shortcomings; for example, the reverse KL tends to be mode-seeking, which reduces output diversity. Recent works have addressed these issues by replacing reverse KL with alternative  $f$ -divergences, such as  $f$ -DPO (Wang et al., 2024) and  $\chi$ PO (Huang et al., 2025). However, these  $f$ -divergence-based constraints still measure policy discrepancy only by comparing token probabilities at identical indices, thereby ignoring semantic relationships between tokens.

To illustrate this limitation, we introduce a simple example in Figure 1. We consider a vocabulary {cat, kitten, dog, table} and compare a reference policy  $\pi_{\text{ref}}$  and two learned policies,  $\pi_1$  and  $\pi_2$ , in the context of next token selection when answering the question “What is in this image?” given a small cat image. In this example,  $\pi_{\text{ref}}$ ,  $\pi_1$ , and  $\pi_2$  assign high probability mass to cat, kitten, and table, respectively. Semantically, (cat, kitten) is more closely related than (cat, table), so we would expect  $\pi_{\text{ref}}$  to be closer to  $\pi_1$  than  $\pi_2$ . However, KL values diverge

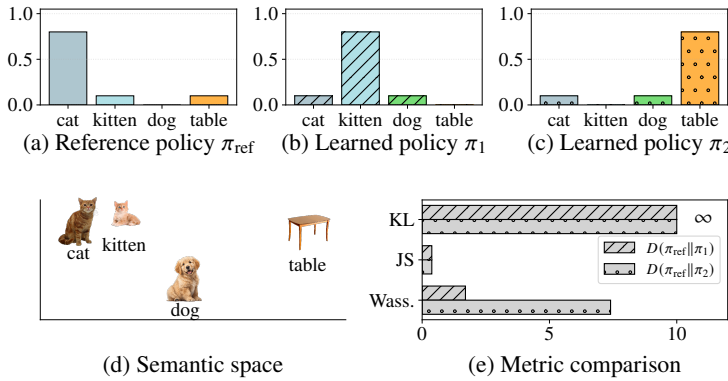


Figure 1: Motivating example for the Wasserstein distance in LLM policy comparison. (a-c) Probability distributions of the reference and learned policies. (d) Semantic space among tokens. (e) Comparison under different divergences, where Wasserstein distance captures semantic relationships that KL and JS divergences fail to reflect.

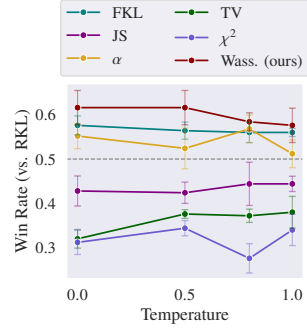


Figure 2: Win rates against KL-based regularization across sampling temperatures on dialogue generation with Gemma-2B, comparing  $f$ -divergences and our Wasserstein distance.

due to the support mismatch, and other  $f$ -divergences such as Jensen-Shannon (JS) divergence assign the same distance to  $\pi_1$  and  $\pi_2$ , failing to reflect semantic proximity.

To overcome this limitation, we introduce a new RLHF regularization framework based on Wasserstein distances, which we refer to as Wasserstein Policy Regularization (WPR). Unlike the KL and other  $f$ -divergences, the Wasserstein metric compares distributions by explicitly considering the geometry of the underlying token space. This enables flexible, user-defined cost functions that naturally encode semantic similarity between tokens. Additionally, it remains well-defined even when the support of two distributions does not overlap. In the context of language modeling, these properties are crucial because policies that assign high probability to semantically related tokens (e.g., *cat* and *kitten*) could likewise be regarded as similar. As illustrated in Figure 1, the Wasserstein distance properly identifies the reference policy  $\pi_{\text{ref}}$  as being closer to  $\pi_1$  than to  $\pi_2$ , thereby capturing semantic proximity that KL and other  $f$ -divergences fail to reflect. As a result, as shown in Figure 2, the policy regularization with the Wasserstein distance achieves superior generation performance compared to KL and other  $f$ -divergence-based approaches, with experimental details provided in Section 5.2.

Building on these properties, we propose a tractable optimization framework that leverages the entropy-regularized Wasserstein distance, i.e., Sinkhorn distance (Cuturi, 2013), as a semantic-aware policy regularizer. Computing this distance requires solving an entropic optimal transport problem; we recast it in the dual and show that the resulting dual variables represent the regularization penalty. This penalty can be incorporated into the reward as token-wise adjustments, analogous to standard KL-based regularization, making the formulation compatible with standard RL algorithms such as PPO (Schulman et al., 2017). The optimal dual variables can be obtained efficiently via the Sinkhorn algorithm with modest overhead. Empirically, our approach outperforms KL- and  $f$ -divergence-based baselines, highlighting the effectiveness of semantic-aware policy distances for RLHF.

## 2 RELATED WORKS

**Aligning Large Language Models** Traditional supervised fine-tuning (SFT) methods have been effective in language generation but shows limitations in aligning outputs with human preferences, such as sentiment (Maas et al., 2011), helpfulness (Askell et al., 2021), harmlessness (Gehman et al., 2020), and truthfulness (Lin et al., 2022). RLHF has become the standard approach for preference alignment (Stiennon et al., 2020; Ouyang et al., 2022). It trains a reward model from human preference data and uses it to optimize the policy via reinforcement learning to better match human preferences (Christiano et al., 2017; Ziegler et al., 2019; Böhm et al., 2019). This approach has enabled successful LLMs such as ChatGPT (Achiam et al., 2023). Recent alternatives avoid explicit reward models, including RAFT (Dong et al., 2023), RRHF (Yuan et al., 2023), and DPO (Rafailov et al., 2023), which reformulate preference alignment as direct policy optimization.

**Regularization for Policy Learning** Methods such as RLHF and DPO incorporate regularization by a reverse KL divergence during preference alignment to prevent the learned policy from deviating significantly from a reference model trained via SFT. While this constrains learning to remain close to the behavior of the reference model, the mode-seeking nature of reverse KL tends to limit output diversity (Wiher et al., 2022; Khalifa et al., 2021; Perez et al., 2022; Glaese et al., 2022). To address this limitation, studies such as  $f$ -DPO (Wang et al., 2024) and  $\chi$ PO (Huang et al., 2025) have been proposed. In parallel, other works (Han et al., 2025; Kim et al., 2025) explore alternative divergences for directly matching the optimal policy, though our focus in this work is on regularization. However,  $f$ -divergence-based methods share a key limitation: they measure the distributional discrepancy solely based on probability values at identical indices, without reflecting the semantic relationships between tokens. In contrast, we propose a novel approach that leverages distance metrics from the Integral Probability Metric (IPM) (Müller, 1997), such as Wasserstein distance (Adler & Lunz, 2018; Panaretos & Zemel, 2019), to enable semantic-aware policy regularization.

**Application of Wasserstein Distance** The Wasserstein distance and its variants, such as the Sinkhorn distance, have been widely applied across many machine learning domains, including generative modeling, robust optimization, and reinforcement learning (Arjovsky et al., 2017; Sinha et al., 2018; Moskovitz et al., 2021; Song et al., 2023; Cui et al., 2024). For example, in generative modeling, Wasserstein GANs (Arjovsky et al., 2017) leverage the Wasserstein distance between the generator distribution and the data distribution to improve training stability and mitigate mode collapse. In robust optimization, adversarial training is formulated using Wasserstein balls around the data distribution to provide certified robustness (Sinha et al., 2018). In reinforcement learning, the Wasserstein natural gradient aligns policy updates with the local optimal-transport geometry in behavioral policy optimization (Moskovitz et al., 2021). Song et al. (2023) explore trust-region policy optimization based on Wasserstein and Sinkhorn distance. Building on this line of work, we explore Wasserstein regularization for RLHF, enabling semantic-aware policy alignment.

### 3 PRELIMINARY

#### 3.1 WASSERSTEIN DISTANCE

The Wasserstein distance between two distributions  $\pi$  and  $\pi'$  is defined as

$$D_W(\pi||\pi') := \min_{P \in U(\pi, \pi')} \mathbb{E}_{(y, y') \sim P} [c(y, y')] = \min_{P \in U(\pi, \pi')} \langle P, C \rangle, \quad (1)$$

where  $U(\pi, \pi') := \{P \in \mathbb{R}_+^{d \times d} | P\mathbf{1}_d = \pi, P^\top \mathbf{1}_d = \pi'\}$  is the set of couplings between  $\pi$  and  $\pi'$ ,  $C \in \mathbb{R}_+^{d \times d}$  is the cost matrix with entries  $C_{y, y'} := c(y, y') \geq 0$ ,  $\langle \cdot, \cdot \rangle$  denotes the Frobenius inner product, and  $d$  is the cardinality of the outcome space.

To obtain a smooth and computationally tractable approximation to the Wasserstein distance, an entropy regularization term is added to the optimal transport objective, yielding the entropy-regularized Wasserstein distance, also known as the Sinkhorn distance (Cuturi, 2013):

$$D_{\tilde{W}}(\pi||\pi') := \min_{P \in U(\pi, \pi')} \left\{ \langle P, C \rangle - \frac{1}{\lambda} \mathcal{H}(P) \right\}, \quad (2)$$

where  $\lambda$  is an entropy regularization hyperparameter, and  $\mathcal{H}(P) := -\sum_{i=1}^d \sum_{j=1}^d P_{ij} (\log P_{ij} - 1)$  is the entropy regularization term, equivalent to the Shannon entropy up to an additive constant.

While the Wasserstein distance directly relies on the Kantorovich dual formulation of optimal transport, the Sinkhorn distance arises from the dual of its entropically regularized variant (Villani, 2008; Peyré et al., 2019):

$$D_W(\pi||\pi') = \max_{\phi, \psi} \left\{ \sum_{i=1}^d \phi_i \pi_i + \sum_{j=1}^d \psi_j \pi'_j \mid \phi_i + \psi_j \leq C_{ij} \forall i, j \right\}, \quad (3)$$

$$D_{\tilde{W}}(\pi||\pi') = \max_{\phi, \psi} \left\{ \sum_{i=1}^d \phi_i \pi_i + \sum_{j=1}^d \psi_j \pi'_j - \frac{1}{\lambda} \sum_{i=1}^d \sum_{j=1}^d \exp(\lambda(\phi_i + \psi_j - C_{ij})) \right\}, \quad (4)$$

where  $\phi$  and  $\psi$  are the dual variables. In the Wasserstein case with the Euclidean cost, the dual variables reduce to a single 1-Lipschitz function, which is typically parameterized by a critic network and optimized with gradient-based methods (Arjovsky et al., 2017). In contrast, the entropy-regularized formulation yields dual optimality conditions corresponding to matrix scaling factors, which can be computed efficiently by the Sinkhorn-Knopp algorithm (Sinkhorn & Knopp, 1967) as closed-form iterations alternating between row and column normalization (Cuturi, 2013; Cuturi & Doucet, 2014).

The entropy-regularized Wasserstein distance produces smoother and denser couplings between distributions, and it converges to the Wasserstein distance as  $\lambda \rightarrow \infty$ . Moreover, compared to the unregularized Wasserstein distance, the Sinkhorn distance can be computed more efficiently, incurring substantially less computational overhead. Since our setting requires computing next-token predictive distributions conditioned on various prompts and partial responses, we employ the entropic regularization variant rather than the critic-based Wasserstein distance, as the former admits the closed-form iterations.<sup>1</sup>

### 3.2 REINFORCEMENT LEARNING FROM HUMAN PREFERENCES (RLHF)

Our goal is to align an autoregressive LLM, denoted as  $\pi_\theta(\mathbf{y}|\mathbf{x})$  where  $\mathbf{x}$  is a user prompt and  $\mathbf{y}$  is a response, with human preferences through reinforcement learning (RL). The RLHF procedure consists of three main stages. First, we perform supervised fine-tuning (SFT) to obtain a reference model  $\pi_{\text{ref}}$ , which serves as the initial aligned model. Second, we train a reward model  $r(\mathbf{x}, \mathbf{y})$  on a preference dataset, enabling the estimation of scalar rewards for responses  $\mathbf{y}$  given prompts  $\mathbf{x}$ . Finally, using both the reference model  $\pi_{\text{ref}}$  and the reward model  $r$ , we optimize the following objective to fine-tune the language model  $\pi_\theta$ :

$$\max_{\pi_\theta} \mathcal{J}(\pi_\theta; \pi_{\text{ref}}) := \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} [\mathbb{E}_{\mathbf{y} \sim \pi_\theta(\mathbf{y}|\mathbf{x})} [r(\mathbf{x}, \mathbf{y})] - \beta D(\pi_\theta(\mathbf{y}|\mathbf{x}) || \pi_{\text{ref}}(\mathbf{y}|\mathbf{x}))], \quad (5)$$

where  $D$  denotes a policy divergence,  $\beta$  is a policy regularization hyperparameter, and  $\mathcal{D}$  is the prompt dataset. This objective encourages the policy to generate responses that maximize reward while remaining close to the reference model.

In most of the previous works, the divergence  $D$  is instantiated as the (reverse) KL divergence:

$$\max_{\pi_\theta} \mathcal{J}_{\text{KL}}(\pi_\theta; \pi_{\text{ref}}) := \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} [\mathbb{E}_{\mathbf{y} \sim \pi_\theta(\mathbf{y}|\mathbf{x})} [r(\mathbf{x}, \mathbf{y})] - \beta D_{\text{KL}}(\pi_\theta(\mathbf{y}|\mathbf{x}) || \pi_{\text{ref}}(\mathbf{y}|\mathbf{x}))] \quad (6)$$

$$= \mathbb{E}_{\mathbf{x}} \left[ \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_\theta(y_n|\mathbf{x}, \mathbf{y}_{1:n-1})} [R(\mathbf{x}, \mathbf{y}_{1:n})] - \beta \sum_{n=1}^N D_{\text{KL}}(\pi_\theta(y_n|\mathbf{x}, \mathbf{y}_{1:n-1}) || \pi_{\text{ref}}(y_n|\mathbf{x}, \mathbf{y}_{1:n-1})) \right] \quad (7)$$

$$= \mathbb{E}_{\mathbf{x}} \left[ \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_\theta(y_n|\mathbf{x}, \mathbf{y}_{1:n-1})} \left[ R(\mathbf{x}, \mathbf{y}_{1:n}) - \beta \log \frac{\pi_\theta(y_n|\mathbf{x}, \mathbf{y}_{1:n-1})}{\pi_{\text{ref}}(y_n|\mathbf{x}, \mathbf{y}_{1:n-1})} \right] \right], \quad (8)$$

where  $D_{\text{KL}}(\pi(y_n) || \pi'(y_n)) := \mathbb{E}_{y_n \sim \pi(y_n)} \left[ \log \frac{\pi(y_n)}{\pi'(y_n)} \right]$ ,  $N$  is the sequence length of  $\mathbf{y}$ , and  $R(\mathbf{x}, \mathbf{y}_{1:n}) = r(\mathbf{x}, \mathbf{y}_{1:N})$  for  $n = N$ , and 0 otherwise. As shown in Eq. (8), the KL regularization term  $D_{\text{KL}}$  can be rewritten as the expectation of the log-ratio between the two policies, which allows standard RL algorithms such as PPO to be applied for optimization. Furthermore, several works (Wang et al., 2024; Huang et al., 2025) have generalized the KL divergence to other  $f$ -divergences and developed tractable optimization formulations accordingly.

However, as mentioned in the Introduction, KL or other  $f$ -divergence measures compare policies solely by token-level probability differences at identical indices, without accounting for the underlying semantic structure of tokens. This limitation prevents them from fully capturing meaningful distributional differences in language generation. To address this, we replace the divergence term with the Wasserstein distance, more precisely the Sinkhorn distance, which naturally incorporates semantic information, and we develop a tractable optimization framework for this objective.

<sup>1</sup>In preliminary experiments, we explored the critic-based Wasserstein distance but found that the resulting policy regularization was insufficient, leading to suboptimal performance.

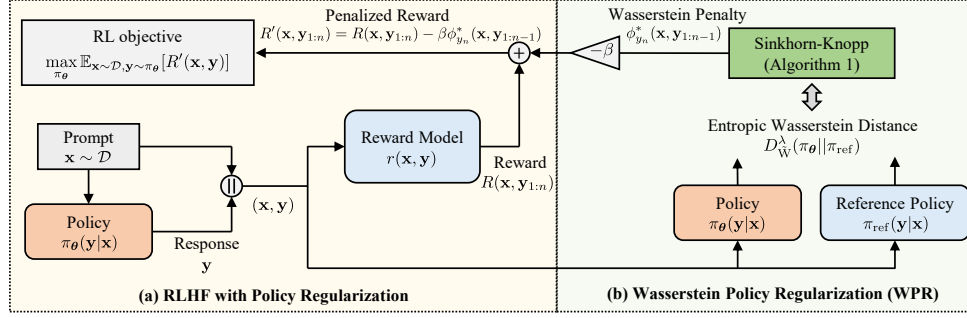


Figure 3: Overview of RLHF with Wasserstein Policy Regularization. (a) Standard RLHF with a policy regularization penalty. (b) Our proposed Wasserstein policy regularization, where the penalty is computed from the optimal dual variables obtained via the Sinkhorn-Knopp algorithm.

## 4 METHOD: WASSERSTEIN POLICY REGULARIZATION

### 4.1 RLHF OBJECTIVE WITH WASSERSTEIN POLICY REGULARIZATION

In this section, we propose Wasserstein Policy Regularization (WPR), which regularizes LLM policies in RLHF using the Wasserstein distance as the statistical distance between policies. We formulate the Wasserstein-regularized RLHF objective by replacing the token-level KL divergence regularization term in Eq. (7) with a Wasserstein regularization term:

$$\max_{\pi_\theta} \mathcal{J}_{\text{W}}(\pi_\theta; \pi_{\text{ref}}) := \quad (9)$$

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[ \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_\theta(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})} [R(\mathbf{x}, \mathbf{y}_{1:n})] - \beta \sum_{n=1}^N D_{\text{W}}(\pi_\theta(y_n | \mathbf{x}, \mathbf{y}_{1:n-1}) || \pi_{\text{ref}}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})) \right]$$

Here,  $D_{\text{W}}(\pi_\theta(y_n | \mathbf{x}, \mathbf{y}_{1:n-1}) || \pi_{\text{ref}}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})) := \min_{\mathbf{P}^{(n)} \sim U_n(\pi_\theta, \pi_{\text{ref}})} \langle \mathbf{P}^{(n)}, \mathbf{C} \rangle$  where  $U_n(\pi_\theta, \pi_{\text{ref}}) := \left\{ \mathbf{P}^{(n)} \in \mathbb{R}_+^{d \times d} | \mathbf{P}^{(n)} \mathbf{1}_d = \pi_\theta(\cdot | \mathbf{x}, \mathbf{y}_{1:n-1}), \mathbf{P}^{(n)\top} \mathbf{1}_d = \pi_{\text{ref}}(\cdot | \mathbf{x}, \mathbf{y}_{1:n-1}) \right\}$ ,  $\mathbf{C} \in \mathbb{R}_+^{d \times d}$  is the cost matrix,  $d$  is the token dictionary size, and  $N$  is the sequence length of  $\mathbf{y}$ . It should be noted that  $\mathbf{P}^{(n)}$  depends on  $(\mathbf{x}, \mathbf{y}_{1:n-1})$ , but we omit these input terms for simplicity.

The next step is to formulate the Wasserstein distance between two token-level discrete distributions so as to obtain a tractable optimization objective for  $\pi_\theta$ . However, computing the exact Wasserstein distance requires solving a linear program, which quickly becomes intractable when the distributional support is large (Kuhn et al., 2019). As discussed in Section 3.1, a widely used approximation is to introduce entropic regularization into the transport problem, referred to as the entropy-regularized Wasserstein distance or Sinkhorn distance (Cuturi, 2013):

$$\max_{\pi_\theta} \mathcal{J}_{\text{W}}^\lambda(\pi_\theta; \pi_{\text{ref}}) := \quad (10)$$

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[ \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_\theta(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})} [R(\mathbf{x}, \mathbf{y}_{1:n})] - \beta \sum_{n=1}^N D_{\text{W}}^\lambda(\pi_\theta(y_n | \mathbf{x}, \mathbf{y}_{1:n-1}) || \pi_{\text{ref}}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})) \right]$$

$$\text{where } D_{\text{W}}^\lambda(\pi_\theta(\cdot | \mathbf{x}, \mathbf{y}_{1:n-1}) || \pi_{\text{ref}}(\cdot | \mathbf{x}, \mathbf{y}_{1:n-1})) := \min_{\mathbf{P}^{(n)} \in U_n} \left\{ \langle \mathbf{P}^{(n)}, \mathbf{C} \rangle - \frac{1}{\lambda} \mathcal{H}(\mathbf{P}^{(n)}) \right\}. \quad (11)$$

We refer to this objective,  $\mathcal{J}_{\text{W}}^\lambda(\pi_\theta; \pi_{\text{ref}})$ , as the entropic Wasserstein-regularized RLHF objective. We now derive the dual problem from the regularized primal transportation problem in Eq. (11). Specifically, we introduce the Lagrangian function  $\mathcal{L}$  corresponding to Eq. (11).

$$\begin{aligned} \mathcal{L}(\mathbf{P}^{(n)}, \phi, \psi) := & \sum_{i=1}^d \sum_{j=1}^d \left( P_{ij}^{(n)} C_{ij} + \frac{1}{\lambda} P_{ij}^{(n)} (\log P_{ij}^{(n)} - 1) \right) \\ & + \sum_{i=1}^d \phi_i ([\pi_\theta]_i - \sum_{k=1}^d P_{ik}^{(n)}) + \sum_{j=1}^d \psi_j ([\pi_{\text{ref}}]_j - \sum_{k=1}^d P_{kj}^{(n)}), \end{aligned} \quad (12)$$

where  $\{\phi_i\}_{i=1}^d$  and  $\{\psi_j\}_{j=1}^d$  are the Lagrange multipliers, introduced to enforce the marginal constraints in  $U_n$ ; specifically, they ensure that the row sums of  $\mathbf{P}^{(n)}$  match  $\pi_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{1:n-1})$  and the column sums match  $\pi_{\text{ref}}(\cdot \mid \mathbf{x}, \mathbf{y}_{1:n-1})$ . Similar to  $\mathbf{P}^{(n)}$ , the dual variables  $\phi$  and  $\psi$  are functions of  $(\mathbf{x}, \mathbf{y}_{1:n-1})$ , but we omit their input terms for brevity unless this causes ambiguity. Based on this Lagrangian, the corresponding dual problem is given by

$$\max_{\phi, \psi} \sum_{i=1}^d \phi_i [\pi_\theta]_i + \sum_{j=1}^d \psi_j [\pi_{\text{ref}}]_j - \sum_{i=1}^d \sum_{j=1}^d \frac{1}{\lambda} \exp(\lambda(\phi_i + \psi_j - C_{ij})), \quad (13)$$

which is derived in Appendix A.1. With strong duality and formulation of the primal solution, we can find the optimal solutions by Proposition 1 (Cuturi & Doucet, 2014).

**Proposition 1.** (Cuturi & Doucet, 2014) *There exists a pair of vectors  $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}_+^d \times \mathbb{R}_+^d$  such that the optimal solutions of  $\mathbf{P}^{(n)}$ ,  $\phi$ , and  $\psi$  are respectively given by*

$$\mathbf{P}^{(n)*} = \text{diag}(\mathbf{u}) \exp(-\lambda \mathbf{C}) \text{diag}(\mathbf{v}), \quad \phi^* = -\frac{1}{\lambda} \log(\mathbf{u}), \quad \psi^* = -\frac{1}{\lambda} \log(\mathbf{v}). \quad (14)$$

We present the proof in Appendix A.2. Note that in our formulation,  $\exp$  denotes the element-wise exponential applied to each entry of  $\mathbf{C}$ . In addition, for any real value  $t$ , the pair of dual variables,  $(\phi + t\mathbf{1}_d, \psi - t\mathbf{1}_d)$  yields the same dual objective value. Hence, the dual optimal solutions are not unique but are determined only up to an additive constant. However, as shown in Theorem 2, when formulating the policy optimization problem based on this optimal solution, the additive term remains constant with respect to the policy and can therefore be ignored, yielding an equivalent problem.

By strong duality, substituting the optimal primal and dual variables obtained in Proposition 1 into the objective in Eq. (13) yields an expression of the entropy-regularized Wasserstein distance in terms of the optimal variables. Plugging this result back into the RLHF formulation in Eq. (10), we obtain a tractable optimization problem, as stated in Theorem 2.

**Theorem 2.** *Let  $\phi^*(\mathbf{x}, \mathbf{y}_{1:n-1})$  denote the optimal dual variables of the entropic optimal transport problem in Eq. (13). Then, the entropic Wasserstein-regularized RLHF in Eq. (10) can be equivalently written as a reward maximization problem with an additional penalty, induced by  $\phi^*$ , i.e.,*

$$\mathcal{J}_{\tilde{W}}(\pi_\theta; \pi_{\text{ref}}) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[ \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_\theta(y_n \mid \mathbf{x}, \mathbf{y}_{1:n-1})} [R(\mathbf{x}, \mathbf{y}_{1:n}) - \beta \phi_{y_n}^*(\mathbf{x}, \mathbf{y}_{1:n-1})] \right] + \mathcal{C}, \quad (15)$$

where  $\mathcal{C}$  is a constant with respect to  $\pi_\theta$ .

The proof is provided in Appendix A.3. Since the objective  $\mathcal{J}_{\tilde{W}}$  of Eq. (15) can be expressed as the sum of token-wise rewards over sampled response sequences, the entropic Wasserstein-regularized RLHF problem, Eq. (10), can be optimized using standard RL methods such as PPO (Schulman et al., 2017). The full RLHF training algorithm is provided in Algorithm 2 of Appendix B.

## 4.2 COMPUTATION OF WASSERSTEIN PENALTY

As shown in the objective of Eq. (15), computing the Wasserstein penalty requires obtaining the optimal dual solution  $\phi$  of the entropic optimal transport problem. To this end, we need to compute the vectors  $\mathbf{u}$  and  $\mathbf{v}$  introduced in Proposition 1. These can be efficiently obtained by applying the Sinkhorn-Knopp algorithm (Sinkhorn & Knopp, 1967) for the matrix scaling problem, as described in Algorithm 1 (Cuturi, 2013; Cuturi & Doucet, 2014).

Specifically, as shown in Proposition 1, the optimal primal solution,  $\mathbf{P}^{(n)*}$ , can be expressed as the product of the positive matrix,  $\exp(-\lambda \mathbf{C})$ , and two diagonal matrices,  $\text{diag}(\mathbf{u})$  and  $\text{diag}(\mathbf{v})$ , with positive entries. Since  $\mathbf{P}^{(n)*}$  is a transportation map, it must be doubly stochastic. Consequently, solving for  $\mathbf{u}$  and  $\mathbf{v}$  in Proposition 1 reduces to a matrix

---

### Algorithm 1 Computation of Wasserstein Penalty via Sinkhorn-Knopp Algorithm

---

**Input:** Learned policy  $\pi_\theta(\cdot \mid \mathbf{x}, \mathbf{y}_{1:n-1})$ , Reference policy  $\pi_{\text{ref}}(\cdot \mid \mathbf{x}, \mathbf{y}_{1:n-1})$ , Cost  $\mathbf{C}$

- 1:  $\mathbf{u} \leftarrow \mathbf{1}_d, \mathbf{v} \leftarrow \mathbf{1}_d, \mathbf{K} \leftarrow \exp(-\lambda \mathbf{C})$
- 2: **while** converged **do**
- 3:    $\text{diag}(\mathbf{u}) \leftarrow \pi_\theta \cdot / \mathbf{K}(\text{diag}(\mathbf{v}))$
- 4:    $\text{diag}(\mathbf{v}) \leftarrow \pi_{\text{ref}} \cdot / \mathbf{K}^\top(\text{diag}(\mathbf{u}))$
- 5: **end while**
- 6:  $\phi \leftarrow -\frac{1}{\lambda} \log(\mathbf{u})$

**Output:** Dual variable  $\phi$

---



scaling problem, which can be solved using the Sinkhorn-Knopp algorithm. This algorithm iteratively rescales the rows and columns of  $\mathbf{P}^{(n)*}$  to match the target marginals  $\pi_{\theta}$  and  $\pi_{\text{ref}}$ , respectively (lines 3-4 in Algorithm 1, where  $\cdot /$  denotes element-wise division).

**Practical Consideration** In practice, the Sinkhorn-Knopp algorithm can be directly applied, but it requires iterative matrix multiplications with the exponential of the cost matrix  $\mathbf{K} := \exp(-\lambda \mathbf{C}) \in \mathbb{R}_+^{d \times d}$ . This incurs  $\mathcal{O}(d^2)$  computational complexity with respect to the dictionary size  $d$ , leading to increased time and memory consumption. To mitigate this, we employ two forms of truncation.

First, during pre-computation of the cost matrix, we apply a *nearest- $k_1$  truncation*. For each token, distances are computed only to its  $k_1$  nearest neighbors. Entries outside this neighborhood are set to zero in  $\mathbf{K}$ , which is equivalent to assigning infinite distance. This yields a sparse  $\mathbf{K}$ , enabling efficient sparse matrix multiplications that reduce both storage and computation. Second, during the Sinkhorn-Knopp algorithm, we apply a *top- $k_2$  truncation*. The distributions  $\pi_{\theta}$  and  $\pi_{\text{ref}}$  are truncated to their top- $k_2$  indices together with the actually sampled index, while the remaining probability mass is aggregated into a dummy index. This reduces the effective support size from  $d$  to at most  $2k_2 + 2$ , lowering the complexity from  $\mathcal{O}(d^2)$  to  $\mathcal{O}(k_2^2)$ . See Appendix C.3 for details of both truncations. Together, these truncations substantially reduce the computational cost of the entropic Wasserstein distance, with training time per step increasing by only 2.5% compared to standard KL regularization.

## 5 EXPERIMENTS

### 5.1 EXPERIMENTAL SETTINGS

**Tasks and Training Details** To evaluate our Wasserstein policy regularization, we conduct open-ended text generation experiments on two datasets: the TL;DR dataset (Völske et al., 2017) for text summarization and the Anthropic Helpful and Harmless (HH-RLHF) dataset (Bai et al., 2022) for dialogue generation. We follow the experimental setup of Chai et al. (2025)<sup>2</sup>, which provides open-source implementations for RLHF research. Our base model is the pre-trained Gemma-2B (Team et al., 2024b), and we use identical training configurations across all baselines and our method, varying only the regularization hyperparameters. For each method, the policy regularization hyperparameter  $\beta$  is selected via grid search to identify the value at which training remained stable, and the best-performing model is reported. For Wasserstein policy regularization, we define the cost function as the Euclidean distance in the fixed token embedding space from the reference policy, set  $\lambda = 100$ , and apply truncation hyperparameters  $k_1 = 512$  and  $k_2 = 128$ . Further experimental details are provided in Appendix C.

**Baselines** We compare regularization based on various divergences with the proposed entropic Wasserstein-based regularization. Specifically, in addition to our approach using the entropic Wasserstein distance in Eq. (10), we evaluate reverse KL (RKL) divergence in Eq. (8), as well as token-level divergence in Eq. (7) instantiated with alternative  $f$ -divergences, including forward KL (FKL), JS,  $\alpha$ -divergence with  $\alpha = 0.5$ , total variation (TV), and  $\chi^2$  divergence. Each  $f$ -divergence can be expressed in the form of a penalty on the reward through its defining function  $f$ , and the corresponding functions for each divergence are summarized in Table 11 of Appendix C.2.

$$\begin{aligned} & \max_{\pi_{\theta}} \mathcal{J}_f(\pi_{\theta}; \pi_{\text{ref}}) \\ &= \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[ \sum_{n=1}^N \mathbb{E}_{\mathbf{y}_n \sim \pi_{\theta}(\mathbf{y}_n | \mathbf{x}, \mathbf{y}_{1:n-1})} \left[ R(\mathbf{x}, \mathbf{y}_{1:n}) - \beta \frac{\pi_{\text{ref}}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})}{\pi_{\theta}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})} f \left( \frac{\pi_{\theta}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})}{\pi_{\text{ref}}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})} \right) \right] \right]. \end{aligned} \quad (16)$$

**Evaluation** We adopt GPT-4 win rate, a widely used evaluation metric in recent LLM studies (Zheng et al., 2023; Chai et al., 2025), as our primary metric. For evaluation, we randomly sample 50 validation instances and generate model responses, repeating this procedure five times. Then, GPT-4 is asked to perform pairwise comparisons between model outputs and report a win rate. We use the GPT-4 evaluation prompts from Chai et al. (2025), with the full prompt included in Appendix C.4. For TL;DR, we assess relevance, coherence, consistency, and fluency; while for HH-RLHF we focus on helpfulness. To reduce evaluation bias, we randomize the order of the responses.

<sup>2</sup><https://github.com/ernie-research/MA-RLHF>

Table 1: Comparison of win rates for policy regularization with various divergences, compared to SFT and RKL-regularized PPO on the TL;DR and the HH-RLHF datasets with the Gemma-2B model.

Divergence	TL;DR		HH-RLHF	
	vs. SFT	vs. RKL	vs. SFT	vs. RKL
RKL	0.848 $\pm$ 0.021	-	0.828 $\pm$ 0.010	-
FKL	0.316 $\pm$ 0.026	0.040 $\pm$ 0.011	0.808 $\pm$ 0.048	0.564 $\pm$ 0.019
JS	0.540 $\pm$ 0.024	0.204 $\pm$ 0.029	0.744 $\pm$ 0.031	0.424 $\pm$ 0.024
$\alpha$ ( $\alpha = 0.5$ )	0.724 $\pm$ 0.031	0.304 $\pm$ 0.016	0.792 $\pm$ 0.047	0.524 $\pm$ 0.046
TV	0.364 $\pm$ 0.039	0.052 $\pm$ 0.021	0.748 $\pm$ 0.038	0.376 $\pm$ 0.010
$\chi^2$	0.904 $\pm$ 0.026	0.540 $\pm$ 0.030	0.796 $\pm$ 0.039	0.344 $\pm$ 0.017
<b>Wasserstein (ours)</b>	<b>0.924</b> $\pm$ 0.019	<b>0.608</b> $\pm$ 0.029	<b>0.852</b> $\pm$ 0.027	<b>0.616</b> $\pm$ 0.039

Table 2: Win rates on TL;DR with Gemma-7B. ‘-2B’ compares to the 2B models in Table 1, and ‘-7B’ to the 7B baselines.

	vs. SFT-2B	vs. RKL-2B	vs. SFT-7B	vs. RKL-7B
RKL	<b>0.948</b>	0.668	0.912	-
<b>Wasserstein</b>	<b>0.948</b>	<b>0.712</b>	<b>0.924</b>	<b>0.532</b>

Table 3: Win rates on HH-RLHF with Qwen1.5-1.8B-Chat.

	vs. SFT	vs. RKL
RKL	0.716	-
<b>Wasserstein</b>	<b>0.752</b>	<b>0.560</b>

## 5.2 QUANTITATIVE RESULTS

**Comparison to Other Divergences** Table 1 reports the performance of summarization and dialogue generation tasks under different policy regularization methods. Additionally, Figure 2 presents win-rate comparisons against RKL across varying sampling temperatures on the HH-RLHF dataset. As the results demonstrate, our proposed Wasserstein policy regularization method achieves the best results on both datasets. In contrast,  $f$ -divergence-based methods rely on probability ratios between policies, which can produce exploding values and unstable training. This issue is evident in the particularly poor TL;DR results of FKL and TV. By comparison, our method remains well-defined even under support mismatch, enabling stable training and delivering superior performance consistently.

**MT-Bench Results** To further empirically validate our approach, we evaluate model using MT-Bench, a GPT-4-based benchmark shown to strongly correlate with human preference judgments (Zheng et al., 2023). We follow the official implementation<sup>3</sup> and apply single-answer grading to models fine-tuned on HH-RLHF. As shown in Table 4, our method achieves the highest performance among all baselines, indicating that semantic-aware regularization improves broader conversational and instruction following abilities.

**Other LLM backbones** To assess the scalability and generalization of WPR, we evaluate the method on larger and architecturally distinct LLM backbones. Using Gemma-7B on the TL;DR summarization task, Table 2 shows that WPR continues to outperform the RKL-regularized baseline. We further demonstrate generalization by training Qwen-1.5-1.8B-Chat (Bai et al., 2023) on HH-RLHF. As shown in Table 3, WPR improves performance over the RKL-regularized model. These results indicate that WPR remains consistently effective across different architectures and model scales.

**Code Generation** We also examine WPR in a different application domain, code generation. Following the experimental settings used in Chai et al. (2025), we assess CodeGemma-7B (Team et al., 2024a) on the APPS dataset (Hendrycks et al., 2021). We compute a reward using the compiler-execution signal employed in prior work (Chai et al., 2025), and we report both the reward and the pass@1 metric over the full 5k test set. As shown in Table 5, WPR achieves consistent improvements across the Introductory, Interview, and Competition levels, as well as in the overall performance.

Table 4: MT-Bench score comparison on Gemma-2B trained on HH-RLHF, evaluated with GPT-4 single-answer grading.

Divergence	Score
RKL	4.000
FKL	4.247
JS	4.197
$\alpha$ ( $\alpha = 0.5$ )	4.256
TV	4.072
$\chi^2$	4.144
<b>Wasserstein</b>	<b>4.272</b>

<sup>3</sup>[https://github.com/lm-sys/FastChat/tree/main/fastchat/llm\\_judge](https://github.com/lm-sys/FastChat/tree/main/fastchat/llm_judge)

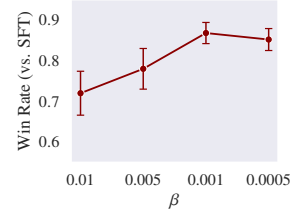


Table 5: Performance comparison on APPS with the CodeGemma-7B model.

	Introductory		Interview		Competition		All	
	Reward	pass@1	Reward	pass@1	Reward	pass@1	Reward	pass@1
SFT	0.1024	23.12	-0.1720	4.86	-0.3239	1.48	-0.1475	7.84
RKL	0.1387	24.00	-0.1316	5.28	-0.2910	1.76	-0.1093	8.32
<b>Wasserstein</b>	<b>0.1606</b>	<b>24.78</b>	<b>-0.1062</b>	<b>5.75</b>	<b>-0.2638</b>	<b>1.92</b>	<b>-0.0843</b>	<b>8.79</b>

Table 6: Ablation study of WPR on TL;DR.

Method		Win rate	
		vs. SFT	vs. RKL
Our default settings		0.924 $\pm$ 0.019	0.608 $\pm$ 0.029
Cost change	(L2 $\rightarrow$ cosine)	0.932 $\pm$ 0.014	<b>0.644</b> $\pm$ 0.047
Decreased $k_1$	(512 $\rightarrow$ 256)	0.920 $\pm$ 0.006	0.572 $\pm$ 0.025
Decreased $k_2$	(128 $\rightarrow$ 64)	0.864 $\pm$ 0.015	0.528 $\pm$ 0.032
Decreased $\lambda$	(100 $\rightarrow$ 10)	0.868 $\pm$ 0.024	0.552 $\pm$ 0.010
Decreased Sinkhorn iterations	(10 $\rightarrow$ 5)	0.708 $\pm$ 0.027	0.328 $\pm$ 0.026
Increased Sinkhorn iterations	(10 $\rightarrow$ 30)	0.880 $\pm$ 0.021	0.536 $\pm$ 0.029

Figure 4: Sensitivity analysis of the policy regularization hyperparameter  $\beta$  on HH-RLHF.

### 5.3 ANALYSIS OF WASSERSTEIN POLICY REGULARIZATION

**Ablation Study** We conduct an ablation study to better understand the effect of components in our Wasserstein policy regularization framework, with results summarized in Table 6. Changing the cost function from the Euclidean to cosine distance yields slightly improved results, suggesting that the framework is robust to the choice of token-level cost metric. Decreasing the truncation parameters  $k_1$  and  $k_2$  or the entropy regularization strength  $\lambda$  leads to a slight drop in performance, though our method still consistently outperforms RKL. Smaller  $k_1$  and  $k_2$  introduce approximation errors in the distance computation, and a smaller  $\lambda$  produces overly sharp couplings that reduce stability. In practice, we use the default settings, which provide consistently robust performance across datasets and configurations. The number of Sinkhorn iterations also affects the distance computation. Reducing iterations from 10 to 5 leads to a sharp drop in performance due to insufficient convergence, while increasing iterations to 30 provides no additional benefit over the default setting. These results suggest that a moderate number of iterations is sufficient for achieving a balance between accuracy and computational efficiency.

Additionally, we analyze the sensitivity analysis of the policy regularization coefficient  $\beta$  in Figure 4. Our Wasserstein-regularized approach achieves stable performance across a broad range of  $\beta$  values, consistently outperforming the SFT baseline. In contrast, we observe that the  $f$ -divergence regularized RLHF yields stable training only within narrow ranges of  $\beta$ , as also reported in the previous work (Wang et al., 2024). While our method demonstrates robustness over a wider range of  $\beta$ , it still requires selecting an appropriate  $\beta$ , highlighting a fundamental limitation of RLHF. Developing approaches that reduce or remove this dependence is an important direction for future work.

**Wasserstein Penalty** Figure 5 compares the KL and Wasserstein penalties computed during training on TL;DR, where both are scaled by the optimal regularization coefficient  $\beta$  and jointly normalized to the range  $[0, 1]$  using a shared min-max range. Note that larger penalties correspond to greater deviation from the reference policy. As shown in the figure, the two penalties exhibit a strong positive correlation, with a Pearson correlation coefficient of 0.917. This result demonstrates that our Wasserstein penalty, similar to the KL penalty, increases as the learned policy differs from the reference policy. Moreover, the fitted trend line has a slope of 0.579, which is less than 1, with a substantial fraction of points lying below the line, indicating that the Wasserstein penalty tends to be more lenient than KL.

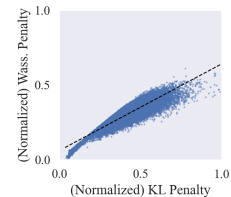


Figure 5: Normalized KL vs. Wasserstein penalty.

Table 7: Pearson correlation between each negative penalty and BERTScore (Zhang et al., 2020).

	TL;DR	HH-RLHF
KL penalty	0.1734	0.0172
<b>Wasserstein penalty</b>	<b>0.2160</b>	<b>0.1749</b>

Table 8: Semantic coherence of top-10 token candidates on each dataset.

	TL;DR	HH-RLHF
RKL	$3.781 \pm 0.005$	$3.690 \pm 0.004$
<b>Wasserstein</b>	<b><math>3.593 \pm 0.003</math></b>	<b><math>3.584 \pm 0.004</math></b>

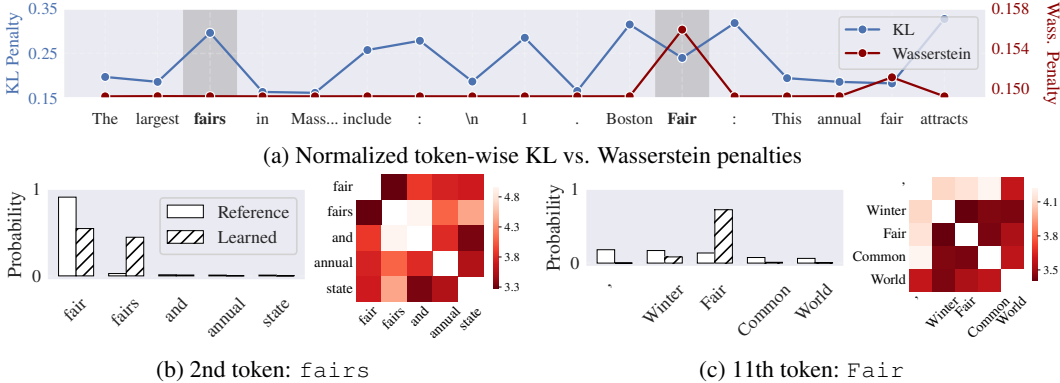


Figure 6: Case study of penalties on Gemma-2B. The prompt is “What fair is the largest fair in Massachusetts?”, and the generated response is “The largest fairs in Massachusetts include: 1. Boston Fair: ...”. (a) Normalized penalties for each generated token. (b-c) Next-token distribution from each policy, along with the relevant cost matrix entries, for the 2nd and 11th tokens highlighted in (a).

To understand how the Wasserstein penalty captures semantic relationships and influences model behavior, we conduct the following analyses. We first evaluate whether the penalty aligns with semantic similarity using BERTScore (Zhang et al., 2020). For responses generated by the reference and learned policies, we compute BERTScore with averaged KL and Wasserstein penalties. Because higher BERTScore indicates greater semantic similarity, we correlate it with the negative value of each penalty. As shown in Table 7, the Wasserstein penalty shows a stronger positive correlation with BERTScore, providing quantitative evidence that WPR better reflects semantic similarity.

We further analyze how the penalty behaves in actual LLM distributions. For the example in Figure 6a, KL fluctuates widely, whereas WPR often assigns minimal penalty. To further investigate, we directly compare the next-token distributions at specific tokens. For semantically similar substitutions in Figure 6b, KL assigns a large penalty due to an exact index mismatch, while WPR gives a small penalty by recognizing semantic proximity. Conversely, when probability mass shifts toward unrelated tokens in Figure 6c, WPR assigns a large penalty, correctly signaling semantic drift.

We also measure semantic coherence of the learned LLMs. For every generated token, we extract the top-10 next-token candidates and compute their mean pairwise embedding distance; smaller distances indicate greater semantic coherence. As shown in Table 8, WPR consistently produces more semantically coherent candidate sets than KL, with statistically significant margins.

Together, these results show that WPR penalizes semantic drift and promotes coherent semantic structure. We conjecture that this semantic awareness contributes to the improved alignment performance.

## 6 CONCLUSION

In this work, we propose a semantic-aware policy regularization framework for RLHF based on the entropy-regularized Wasserstein distance, which captures semantic similarity between tokens beyond the limits of KL and other  $f$ -divergences. By formulating the regularization in the dual space, our method yields tractable penalties compatible with standard RL algorithms, while remaining computationally efficient via the Sinkhorn-Knopp algorithm. Experiments on summarization and dialogue generation tasks demonstrate consistent improvements over KL- and  $f$ -divergence-based baselines, with higher win rates and MT-Bench score. These results highlight the effectiveness of semantic-aware policy distances for stable and robust alignment of large language models.

## ACKNOWLEDGMENTS

This work was supported by the InnoCORE program of the Ministry of Science and ICT (N10260008) (50%). This work was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2026-RS-2024-00437268) (50%).

## REFERENCES

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- Jonas Adler and Sebastian Lunz. Banach wasserstein gan. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.
- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.
- Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- Florian Böhm, Yang Gao, Christian M. Meyer, Ori Shapira, Ido Dagan, and Iryna Gurevych. Better rewards yield better summaries: Learning to summarise without references. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3110–3120, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1307.
- Yekun Chai, Haoran Sun, Huang Fang, Shuohuan Wang, Yu Sun, and Hua Wu. MA-RLHF: Reinforcement learning from human feedback with macro actions. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- Xiao Cui, Yulei Qin, Yuting Gao, Enwei Zhang, Zihan Xu, Tong Wu, Ke Li, Xing Sun, Wengang Zhou, and Houqiang Li. Sinkhorn distance minimization for knowledge distillation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14846–14858, 2024.

- Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pp. 685–693. PMLR, 2014.
- Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, KaShun SHUM, and Tong Zhang. RAFT: Reward ranked finetuning for generative foundation model alignment. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12634–12651. PMLR, 21–27 Jul 2024.
- Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3356–3369, 2020.
- Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*, 2022.
- Jiaqi Han, Mingjian Jiang, Yuxuan Song, Stefano Ermon, and Minkai Xu.  $f$ -po: Generalizing preference optimization with  $f$ -divergence minimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 1144–1152. PMLR, 2025.
- Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring coding challenge competence with APPS. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
- Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D. Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J Foster. Correcting the mythos of KL-regularization: Direct alignment without overoptimization via chi-squared preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Binyuan Hui, Jian Yang, Zeyu Cui, Jiayi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.
- Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A distributional approach to controlled text generation. In *International Conference on Learning Representations*, 2021.
- Yeongmin Kim, Heesun Bae, Byeonghu Na, and Il-Chul Moon. Preference optimization by estimating the ratio of the data distribution. *arXiv preprint arXiv:2505.19601*, 2025.
- Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pp. 130–166. Informa, 2019.
- Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229.
- Jiate Liu, Yiqin Zhu, Kaiwen Xiao, QIANG FU, Xiao Han, Yang Wei, and Deheng Ye. RLTF: Reinforcement learning from unit test feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.
- Ted Moskovitz, Michael Arbel, Ferenc Huszar, and Arthur Gretton. Efficient wasserstein natural gradients for reinforcement learning. In *International Conference on Learning Representations*, 2021.
- Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Victor M Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, 6(1):405–431, 2019.
- Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.225.
- Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.
- John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- Parshin Shojaee, Aneesh Jain, Sindhu Tipirneni, and Chandan K. Reddy. Execution-based code generation using deep reinforcement learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.
- Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.
- Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.
- Jun Song, Niao He, Lijun Ding, and Chaoyue Zhao. Provably convergent policy optimization via metric-aware trust region methods. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.
- CodeGemma Team, Heri Zhao, Jeffrey Hui, Joshua Howland, Nam Nguyen, Siqi Zuo, Andrea Hu, Christopher A Choquette-Choo, Jingyue Shen, Joe Kelley, et al. Codegemma: Open code models based on gemma. *arXiv preprint arXiv:2406.11409*, 2024a.



- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024b.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.
- Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 59–63, 2017.
- Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *The Twelfth International Conference on Learning Representations*, 2024.
- Gian Wiher, Clara Meister, and Ryan Cotterell. On decoding strategies for neural text generators. *Transactions of the Association for Computational Linguistics*, 10:997–1012, 2022.
- Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, Conglong Li, Connor Holmes, et al. DeepSpeed-chat: Easy, fast and affordable rlhf training of chatgpt-like models at all scales. *arXiv preprint arXiv:2308.01320*, 2023.
- Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. RRHF: Rank responses to align language models with human feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

## A PROOFS AND DERIVATIONS

### A.1 DERIVATION OF EQ. (13)

We derive the dual problem in Eq. (13) from the Lagrangian  $\mathcal{L}$  in Eq. (12) constructed for the entropic primal transportation problem in Eq. (11). First, we rewrite the Lagrangian as follows:

$$\mathcal{L}(\mathbf{P}^{(n)}, \phi, \psi) \quad (17)$$

$$\begin{aligned} &:= \sum_{i=1}^d \sum_{j=1}^d \left( P_{ij}^{(n)} C_{ij} + \frac{1}{\lambda} P_{ij}^{(n)} (\log P_{ij}^{(n)} - 1) \right) \\ &\quad + \sum_{i=1}^d \phi_i ([\pi_{\theta}]_i - \sum_{k=1}^d P_{ik}^{(n)}) + \sum_{j=1}^d \psi_j ([\pi_{\text{ref}}]_j - \sum_{k=1}^d P_{kj}^{(n)}) \end{aligned} \quad (18)$$

$$= \sum_{i=1}^d \sum_{j=1}^d P_{ij}^{(n)} \left( C_{ij} - \phi_i - \psi_j + \frac{1}{\lambda} (\log P_{ij}^{(n)} - 1) \right) + \sum_{i=1}^d \phi_i ([\pi_{\theta}]_i) + \sum_{j=1}^d \psi_j ([\pi_{\text{ref}}]_j), \quad (19)$$

where  $\{\phi_i\}_{i=1}^d$  and  $\{\psi_j\}_{j=1}^d$  are the Lagrange multipliers. Based on this Lagrangian, the primal and dual problem can be written as follows:

$$\text{(Primal)} \quad \min_{\mathbf{P}^{(n)}} \max_{\phi, \psi} \mathcal{L}(\mathbf{P}^{(n)}, \phi, \psi), \quad (20)$$

$$\text{(Dual)} \quad \max_{\phi, \psi} \min_{\mathbf{P}^{(n)}} \mathcal{L}(\mathbf{P}^{(n)}, \phi, \psi). \quad (21)$$

By differentiating the Lagrangian with respect to  $P_{ij}^{(n)}$ , we derive the condition that the optimal  $P_{ij}^{(n)*}$  satisfy as follows:

$$C_{ij} - \phi_i - \psi_j + \frac{1}{\lambda} \log P_{ij}^{(n)*} = 0 \quad \Leftrightarrow \quad P_{ij}^{(n)*} = \exp(\lambda(\phi_i + \psi_j - C_{ij})). \quad (22)$$

Therefore, by substituting the optimal  $P_{ij}^{(n)*}$  in Eq. (22) into Eq. (19), we can express the dual problem of Eq. (21) as

$$\max_{\phi, \psi} \sum_{i=1}^d \phi_i [\pi_{\theta}]_i + \sum_{j=1}^d \psi_j [\pi_{\text{ref}}]_j - \sum_{i=1}^d \sum_{j=1}^d \frac{1}{\lambda} \exp(\lambda(\phi_i + \psi_j - C_{ij})). \quad (23)$$

### A.2 PROOF OF PROPOSITION 1

**Proposition 1.** (*Cuturi & Doucet, 2014*) *There exists a pair of vectors  $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}_+^d \times \mathbb{R}_+^d$  such that the optimal solutions of  $\mathbf{P}^{(n)}$ ,  $\phi$ , and  $\psi$  are respectively given by*

$$\mathbf{P}^{(n)*} = \text{diag}(\mathbf{u}) \exp(-\lambda \mathbf{C}) \text{diag}(\mathbf{v}), \quad \phi^* = -\frac{1}{\lambda} \log(\mathbf{u}), \quad \psi^* = -\frac{1}{\lambda} \log(\mathbf{v}). \quad (14)$$

*Proof.* From Eq. (22) in the derivation of Appendix A.1, the optimal  $P_{ij}^{(n)*}$  can be written as

$$P_{ij}^{(n)*} = \exp(\lambda(\phi_i + \psi_j - C_{ij})) = \exp(\lambda\phi_i) \exp(-\lambda C_{ij}) \exp(\lambda\psi_j). \quad (24)$$

Defining the positive kernel  $\mathbf{K} := \exp(-\lambda \mathbf{C})$ , where the exponential is applied element-wise, and the scaling vectors  $\mathbf{u} := \exp(\lambda\phi)$  and  $\mathbf{v} := \exp(\lambda\psi)$ , the optimal coupling admits the compact representation

$$\mathbf{P}^{(n)} = \text{diag}(\mathbf{u}) \mathbf{K} \text{diag}(\mathbf{v}). \quad (25)$$

Since  $\mathbf{K}$  is strictly positive, the Sinkhorn-Knopp theorem (Sinkhorn & Knopp, 1967) guarantees the existence and uniqueness (up to an additive constant) of strictly positive scaling vectors  $\mathbf{u}, \mathbf{v} \in \mathbb{R}_+^d$  such that  $\mathbf{P}^{(n)} \in U_n(\pi_{\theta}, \pi_{\text{ref}})$ . Hence, the primal optimum  $\mathbf{P}^{(n)}$  is unique and necessarily of the form  $\text{diag}(\mathbf{u}) \mathbf{K} \text{diag}(\mathbf{v})$ .

Finally, by the definitions of  $\mathbf{u}$  and  $\mathbf{v}$ , the corresponding optimal dual variables are given by

$$\phi^* = -\frac{1}{\lambda} \log \mathbf{u}, \quad \psi^* = -\frac{1}{\lambda} \log \mathbf{v}, \quad (26)$$

which yields the stated representation

$$\mathbf{P}^{(n)*} = \text{diag}(\mathbf{u}) \exp(-\lambda \mathbf{C}) \text{diag}(\mathbf{v}), \quad \phi^* = -\frac{1}{\lambda} \log \mathbf{u}, \quad \psi^* = -\frac{1}{\lambda} \log \mathbf{v}. \quad (27)$$

□

### A.3 PROOF OF THEOREM 2

**Theorem 2.** Let  $\phi^*(\mathbf{x}, \mathbf{y}_{1:n-1})$  denote the optimal dual variables of the entropic optimal transport problem in Eq. (13). Then, the entropic Wasserstein-regularized RLHF in Eq. (10) can be equivalently written as a reward maximization problem with an additional penalty, induced by  $\phi^*$ , i.e.,

$$\mathcal{J}_{\tilde{W}}(\pi_{\theta}; \pi_{\text{ref}}) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[ \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_{\theta}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})} [R(\mathbf{x}, \mathbf{y}_{1:n}) - \beta \phi_{y_n}^*(\mathbf{x}, \mathbf{y}_{1:n-1})] \right] + \mathcal{C}, \quad (15)$$

where  $\mathcal{C}$  is a constant with respect to  $\pi_{\theta}$ .

*Proof.* First, the objective of Wasserstein-regularized RLHF can be written as

$$\mathcal{J}_{\tilde{W}}(\pi_{\theta}; \pi_{\text{ref}}) := \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[ \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_{\theta}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})} [R(\mathbf{x}, \mathbf{y}_{1:n})] - \beta \sum_{n=1}^N D_{\tilde{W}}^{\lambda}(\pi_{\theta}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1}) || \pi_{\text{ref}}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})) \right]. \quad (28)$$

By strong duality, the entropic Wasserstein distance  $D_{\tilde{W}}^{\lambda}$  is equal to the optimal objective value of the dual problem in Eq. (13). Substituting the optimal solutions from Proposition 1, we obtain

$$\begin{aligned} & D_{\tilde{W}}^{\lambda}(\pi_{\theta}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1}) || \pi_{\text{ref}}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})) \\ &= \sum_{i=1}^d \phi_i^*(\mathbf{x}, \mathbf{y}_{1:n-1}) \pi_{\theta}(y_n = i | \mathbf{x}, \mathbf{y}_{1:n-1}) + \sum_{j=1}^d \psi_j^*(\mathbf{x}, \mathbf{y}_{1:n-1}) \pi_{\text{ref}}(y_n = j | \mathbf{x}, \mathbf{y}_{1:n-1}) \end{aligned} \quad (29)$$

$$- \sum_{i=1}^d \sum_{j=1}^d \frac{1}{\lambda} \exp(\lambda(\phi_i^*(\mathbf{x}, \mathbf{y}_{1:n-1}) + \psi_j^*(\mathbf{x}, \mathbf{y}_{1:n-1}) - C_{ij})) \quad (30)$$

$$= \sum_{i=1}^d \phi_i^*(\mathbf{x}, \mathbf{y}_{1:n-1}) \pi_{\theta}(y_n = i | \mathbf{x}, \mathbf{y}_{1:n-1}) + \mathcal{C} \quad (31)$$

$$= \mathbb{E}_{y_n \sim \pi_{\theta}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})} [\phi_{y_n}^*(\mathbf{x}, \mathbf{y}_{1:n-1})] + \mathcal{C}, \quad (32)$$

where  $\mathcal{C}$  denotes a constant with respect to  $\pi_{\theta}$ .

Substituting Eq. (32) into Eq. (28), the objective reduces to a reward maximization problem with an additional penalty induced by the dual variables  $\phi^*$ :

$$\begin{aligned} & \mathcal{J}_{\tilde{W}}(\pi_{\theta}; \pi_{\text{ref}}) \\ &= \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[ \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_{\theta}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})} [R(\mathbf{x}, \mathbf{y}_{1:n})] - \beta \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_{\theta}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})} [\phi_{y_n}^*(\mathbf{x}, \mathbf{y}_{1:n-1})] + \mathcal{C} \right] \end{aligned} \quad (33)$$

$$= \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[ \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_{\theta}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})} [R(\mathbf{x}, \mathbf{y}_{1:n}) - \beta \phi_{y_n}^*(\mathbf{x}, \mathbf{y}_{1:n-1})] \right] + \mathcal{C} \quad (34)$$

□

## B TRAINING ALGORITHM OF RLHF WITH WASSERSTEIN POLICY REGULARIZATION

---

### Algorithm 2 RLHF with Wasserstein Policy Regularization

---

**Input:** Current policy  $\pi_\theta$ , Old policy  $\pi_{\theta^-}$ , Reference policy  $\pi_{\text{ref}}$ , Reward model  $r(\mathbf{x}, \mathbf{y})$ , Cost matrix  $C$ , Dataset  $\mathcal{D}$

- 1: **for**  $t_{\text{train}} = 1$  **to**  $T_{\text{train}}$  **do**
- 2:   Sample  $\mathbf{y}_{1:d} \sim \pi_\theta(\cdot|\mathbf{x})$  for  $\mathbf{x} \sim \mathcal{D}$  (Computation with Batch Samples)
- 3:   Get  $R(\mathbf{x}, \mathbf{y}_{1:n})$  with reward model  $r$  for  $n = \{1, 2, \dots, d\}$
- 4:   Compute  $\phi_{y_n}^*(\mathbf{x}, \mathbf{y}_{1:n-1})$  via Algorithm 1 using  $C$  for  $n = \{1, 2, \dots, d\}$
- 5:   Obtain  $\hat{A}_n$  for  $n = \{1, 2, \dots, d\}$  via Eq. (35)
- 6:   Compute  $\nabla_\theta \mathcal{J}_{\tilde{W}}(\theta)$  via Eq. (40)
- 7:   Compute  $\nabla_\psi \mathcal{L}_V(\psi)$  via Eq. (38)
- 8:   Update  $\theta \leftarrow \theta + \eta_\pi \nabla_\theta \mathcal{J}_{\tilde{W}}(\theta)$ ,  $\psi \leftarrow \psi - \eta_V \nabla_\psi \mathcal{L}_V(\psi)$  and  $\theta^- \leftarrow \theta$
- 9: **end for**

**Output:** Learned policy  $\pi_\theta$

---

In this section, we present the detailed training algorithm for RLHF with Wasserstein Policy Regularization (WPR). As in conventional RLHF (Ouyang et al., 2022), we iteratively sample response as  $y_n \sim \pi_{\theta^-}(\mathbf{x}, \mathbf{y}_{1:n-1})$  to get  $\mathbf{y}_{1:d}$ . Here,  $\pi_{\theta^-}$  is old policy whose parameters  $\theta^-$  are periodically updated by that of the current policy,  $\pi_\theta$ . At each token generation step  $n$ , we adopt Generalized Advantage Estimation (GAE) (Schulman et al., 2016) for penalized reward  $R'(\mathbf{x}, \mathbf{y}_{1:n}) = R(\mathbf{x}, \mathbf{y}_{1:n}) - \beta \phi_{y_n}^*(\mathbf{x}, \mathbf{y}_{1:n-1})$  in Eq. (15).

Then, the advantage with GAE denoted as  $\hat{A}_n$  at each step  $n$  can be expressed as

$$\hat{A}_n = \sum_{l \geq 0} (\gamma \lambda)^l \delta_{n+l}, \quad (35)$$

where

$$\delta_n = R'(\mathbf{x}, \mathbf{y}_{1:n}) + \gamma V_\psi(\mathbf{x}, \mathbf{y}_{1:n}) - V_\psi(\mathbf{x}, \mathbf{y}_{1:n-1}). \quad (36)$$

Here,  $\gamma$  is a discount factor;  $\lambda$  is a hyperparameter for GAE; and  $V_\psi$  is a value network, which estimates the discounted cumulative reward or return of given state  $(\mathbf{x}, \mathbf{y}_{1:n})$ , denoted as  $\hat{G}_n$ . Thus, the learning loss for  $V_\psi$  is defined as

$$\mathcal{L}_V(\psi) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[ \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_{\theta^-}(y_n|\mathbf{x}, \mathbf{y}_{1:n-1})} \left[ (V_\psi(\mathbf{x}, \mathbf{y}_{1:n}) - \hat{G}_n)^2 \right] \right]. \quad (37)$$

Then, its gradient is expressed as follows.

$$\nabla_\psi \mathcal{L}_V(\psi) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[ \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_{\theta^-}(y_n|\mathbf{x}, \mathbf{y}_{1:n-1})} \left[ 2(V_\psi(\mathbf{x}, \mathbf{y}_{1:n}) - \hat{G}_n) \nabla_\psi V_\psi(\mathbf{x}, \mathbf{y}_{1:n}) \right] \right] \quad (38)$$

In RLHF,  $V_\psi$  and  $\pi_\theta$  are updated together. By substituting a step-wise penalized reward  $R'(\mathbf{x}, \mathbf{y}_{1:n})$  with  $\hat{A}_n$ , Eq. (15) is expressed as

$$\mathcal{J}_{\tilde{W}}(\theta) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[ \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_{\theta^-}(y_n|\mathbf{x}, \mathbf{y}_{1:n-1})} \left[ \frac{\pi_\theta}{\pi_{\theta^-}} \hat{A}_n \right] \right] + C. \quad (39)$$

Here, we denote  $\mathcal{J}_{\tilde{W}}(\pi_\theta; \pi_{\theta^-}, \pi_{\text{ref}})$  as  $\mathcal{J}_{\tilde{W}}(\theta)$  for simplicity and  $\frac{\pi_\theta}{\pi_{\theta^-}}$  is an importance weight. Then, the gradient of  $\mathcal{J}_{\tilde{W}}$  is computed as

$$\nabla_\theta \mathcal{J}_{\tilde{W}}(\theta) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[ \sum_{n=1}^N \mathbb{E}_{y_n \sim \pi_{\theta^-}(y_n|\mathbf{x}, \mathbf{y}_{1:n-1})} \left[ \left( \frac{\pi_\theta}{\pi_{\theta^-}} \hat{A}_n \right) \nabla_\theta \log \pi_\theta \right] \right]. \quad (40)$$

In practice, clipping mechanism for advantage computation (Schulman et al., 2017) is adopted for Eq. (39) and  $V_\psi(\mathbf{x}, \mathbf{y}_{1:n-1})$  in Eq. (38) is also clipped for a stable training. Algorithm 2 presents the overall training framework for RLHF with Wasserstein Policy Regularization. In Algorithm 2,  $t_{\text{train}}$  and  $T_{\text{train}}$  are the training step and the maximum training step, respectively. At  $t_{\text{train}} = 1$ , current policy  $\pi_\theta$ , old policy  $\pi_{\theta^-}$ , and reference policy  $\pi_{\text{ref}}$  are all initialized with SFT model. In Line # 8 in Algorithm 2, we update  $\theta^-$  with updated  $\theta$  at every training step, and  $\eta_\pi$  and  $\eta_V$  are learning rate for  $\pi_\theta$  and  $V_\psi$ , respectively.

## C ADDITIONAL EXPERIMENTAL SETTINGS

### C.1 DATASETS

**TL;DR** For the summarization task, the policy is trained to generate concise summaries of Reddit posts. The dataset includes 93K preference pairs for training and 86K pairs for validation. Training data is derived from the Reddit TL;DR corpus (Völske et al., 2017). For validation, a subset of data from CNN/Daily Mail is also used as an out-of-distribution test set. The dataset is downloaded from Hugging Face.<sup>4</sup>

**HH-RLHF** For dialogue generation, we use the Anthropic HH-RLHF dataset (Bai et al., 2022), where the policy is trained to produce responses that are both helpful and harmless in single-turn and multi-turn dialogue settings. It comprises 112K preference-labeled instances for training and an additional 12.5K instances for validation. The dataset can be downloaded from Hugging Face.<sup>5</sup>

**APPS** For code generation, we use the APPS dataset (Hendrycks et al., 2021), which provides a diverse set of programming problems requiring executable Python solutions. The dataset consists of 5K training instances and 5K validation instances, each containing a natural language problem description paired with unit tests for automated evaluation. The dataset can be downloaded from Hugging Face.<sup>6</sup>

### C.2 MODEL TRAINING DETAILS

We follow the experimental setup of Chai et al. (2025), which provides open-source implementations for RLHF research.<sup>7</sup> This implementation is based on the DeepSpeed-Chat package (Yao et al., 2023), and we adopt its configuration as the default setting. Our base model is the pre-trained Gemma-2B (Team et al., 2024b), and we use identical training configurations across all baselines and our method, varying only the regularization hyperparameters. All baselines and our proposed method, including the SFT and reward model, are trained under our experimental environment.

**Supervised Fine-Tuning (SFT)** We split each dataset into three subsets and allocate 20% for supervised fine-tuning. Prompts are paired with their preferred responses to construct instruction data. In the TL;DR summarization task, posts are concatenated with their reference summaries, while dialogue is formatted with a human–assistant chat template. For this stage, we employ the Gemma-2B model as the backbone. The training configuration specifies a batch size of 512, a learning rate of  $5 \times 10^{-5}$ , a cosine learning rate scheduler with a warmup ratio of 0.1, and a total of 3 epochs.

**Reward Model Training** InstructGPT (Ouyang et al., 2022) mitigates distributional mismatch by fine-tuning the reward model on the same dataset used for SFT. Following this approach, we also train our reward model on the identical dataset. In this stage, 40% of the data is used for reward model training. Preference annotations are processed in the same way as in SFT. The reward model is initialized from the SFT checkpoint. The training configuration specifies a batch size of 64, a learning rate of  $1 \times 10^{-5}$ , a cosine learning rate scheduler with a warmup ratio of 0.1, and a single epoch.

For code generation, we follow prior work (Chai et al., 2025) and use a reward function derived directly from the compiler execution signal, without training an additional reward model. We adopt

<sup>4</sup>[https://huggingface.co/datasets/openai/summarize\\_from\\_feedback](https://huggingface.co/datasets/openai/summarize_from_feedback)

<sup>5</sup><https://huggingface.co/datasets/Dahoas/full-hh-rlhf>

<sup>6</sup><https://huggingface.co/datasets/codeparrot/apps>

<sup>7</sup><https://github.com/ernie-research/MA-RLHF>



Table 9: Hyperparameters for PPO training.

Hyperparameter	Value
PPO epochs	1
Rollout	1
Clip ratio	0.2
$\lambda$ in GAE	0.95
$\gamma$ in GAE	1
Max prompt length	512
Max response length	256
Warmup steps	200
Temperature	0.8
Top-p	1.0
Top-k	50

Table 10: Policy regularization hyperparameter  $\beta$  for each method.

Divergence	TL;DR	HH-RLHF
RKL	0.005	0.001
FKL	0.05	0.0001
JS	0.05	0.01
$\alpha$ ( $\alpha = 0.5$ )	0.01	0.05
TV	0.01	0.01
$\chi^2$	0.001	0.001
<b>Wasserstein</b>	0.05	0.0005

Table 11: Corresponding functions for each  $f$ -divergences.

Divergence	$f(u)$
RKL	$u \log u$
FKL	$-\log u$
JS	$u \log u - (u + 1) \log(\frac{u+1}{2})$
$\alpha$	$\frac{1}{\alpha(\alpha-1)}(u^{1-\alpha} - (1-\alpha)u - \alpha)$
TV	$\frac{1}{2} u - 1 $
$\chi^2$	$(u - 1)^2$

the adaptive compiler-based reward used in previous studies (Chai et al., 2025; Shojaei et al., 2023; Liu et al., 2023). For a generated solution  $\mathbf{y}$  to a problem  $\mathbf{x}$ , the reward is defined as:

$$r(\mathbf{x}, \mathbf{y}) = \begin{cases} -0.3 + 1.3 \cdot \frac{N_{\text{pass}}}{N_{\text{pass}} + N_{\text{fail}}}, & \text{if } y \text{ compiles successfully,} \\ -0.6, & \text{if } y \text{ raises a runtime error,} \\ -1.0, & \text{if } y \text{ fails to compile.} \end{cases}$$

Here,  $N_{\text{pass}}$  and  $N_{\text{fail}}$  denote the number of unit tests passed and failed, respectively.

**Policy Optimization with PPO** The remaining 40% of the dataset is used for PPO training. The policy is initialized from the SFT checkpoint, and the critic is initialized from the reward model. We use the same SFT and reward model checkpoints for all baselines. The training configuration uses a batch size of 256, learning rates of  $1.5 \times 10^{-5}$  for both the policy and the critic, and runs for one epoch. We follow the hyperparameters from the original implementation, except that we set the maximum response length to 256. The hyperparameters are summarized in Table 9.

**Policy Regularization** For each method, the policy regularization hyperparameter  $\beta$  is selected via grid search to identify the value at which training remained stable, and we report the best-performing model. Specifically, we perform a grid search over  $\{0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001\}$ . The resulting  $\beta$  values used for each baseline are summarized in Table 10. Each  $f$ -divergence can be expressed in the form of a penalty on the reward through its defining function  $f$ , and the corresponding functions for each divergence are summarized in Table 11.

$$\begin{aligned} & \max_{\pi_{\theta}} \mathcal{J}_f(\pi_{\theta}; \pi_{\text{ref}}) \\ &= \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \left[ \sum_{n=1}^N \mathbb{E}_{\mathbf{y}_n \sim \pi_{\theta}(\mathbf{y}_n | \mathbf{x}, \mathbf{y}_{1:n-1})} \left[ R(\mathbf{x}, \mathbf{y}_{1:n}) - \beta \frac{\pi_{\text{ref}}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})}{\pi_{\theta}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})} f\left(\frac{\pi_{\theta}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})}{\pi_{\text{ref}}(y_n | \mathbf{x}, \mathbf{y}_{1:n-1})}\right) \right] \right]. \end{aligned} \quad (41)$$

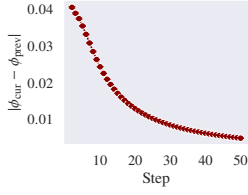


Figure 7: Convergence of the Sinkhorn-Knopp algorithm.

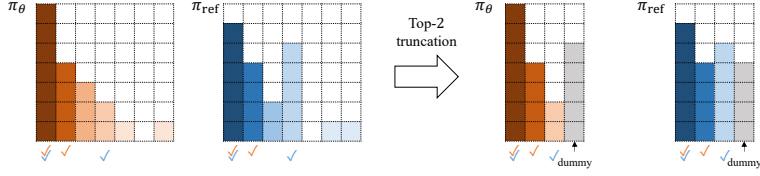


Figure 8: Example of top- $k_2$  truncation where  $k_2 = 2$ .

For WPR, we define the cost function as the Euclidean distance in the fixed token embedding space from the SFT model, set  $\lambda = 100$ , and apply truncation hyperparameters  $k_1 = 512$  and  $k_2 = 128$ . The number of Sinkhorn iterations is set to 10 for TL;DR and 50 for HH-RLHF.

### C.3 SHINKORN ALGORITHM DETAILS

**Stopping Criterion** The Sinkhorn-Knopp iterations involve alternating updates of the scaling vectors  $\mathbf{u}$  and  $\mathbf{v}$ . Since the dual variable  $\phi$  is ultimately used as the regularization penalty, we monitor convergence based on the change in  $\phi = -\frac{1}{\lambda} \log \mathbf{u}$ . Iterations are terminated when the change in  $\phi$  falls below a pre-defined tolerance, which we use  $10^{-4}$ . For practicality, we also impose a maximum number of Sinkhorn iterations, as specified in Appendix C.2.

**Numerical Stability** The Sinkhorn updates involve repeated rescaling operations and log computations when recovering  $\phi$ , which can lead to numerical instabilities. To mitigate this, we add small constants to denominators and log arguments. Figure 7 plots the evolution of the convergence metric across iterations on Gemma-2B experiments with HH-RLHF, demonstrating that the truncated Sinkhorn procedure converges stably in practice.

**Nearest- $k_1$  Truncation for Cost Matrix** Since the cost matrix  $\mathbf{C}$  is fixed throughout training, we pre-compute  $\mathbf{K} = \exp(-\lambda \mathbf{C})$ . However, storing the full  $\mathbf{K}$  is infeasible for vocabularies of extremely large tokens. To address this, for each token we retain only its  $k_1$  nearest neighbors and set all other entries to zero, which is equivalent to assigning infinite cost to distant tokens. We additionally enforce symmetry by mirroring retained entries so that the sparse kernel remains consistent. This sparsification enables efficient sparse-matrix multiplications during Sinkhorn iterations.

**Top- $k_2$  Truncation for Token Distributions** For computational efficiency, we also truncate the token distributions of both the target policy and the reference policy. Specifically, we retain the top- $k_2$  probability indices along with the index of the sampled token. The remaining probability mass is aggregated into a dummy index. Because the sampled token is always included, the required dual variable  $\pi_{y_n}^*$  can be recovered. A conceptual illustration is provided in Figure 8.

### C.4 EVALUATION DETAILS

We adopt GPT-4 win rate, a widely used evaluation metric in recent LLM studies (Zheng et al., 2023; Chai et al., 2025), as our main evaluation measure. For each comparison, we randomly sample 50 validation instances and generate model responses, repeating this procedure five times. Unless otherwise noted, the sampling temperature is fixed at 0.5. Then, GPT-4 is asked to perform pairwise comparisons between model outputs and compute the win rate. We use the gpt-4o-2024-05-13 model for all evaluations. We follow the GPT-4 evaluation prompts provided by Chai et al. (2025), and for completeness, we include the full prompt below. For TL;DR, we assess relevance, coherence, consistency, and fluency; while for HH-RLHF we focus on helpfulness. To reduce evaluation bias, we randomize the order of the responses.

**GPT-4 Evaluation Prompt for TL;DR**

You will be given two summaries written for an article. Your task is to pick the better one between them, based on the four criteria. Please make sure you read and understand these instructions very carefully.

**Relevance** - selection of important content from the source. The summary should include only important information from the source document. Annotators were instructed to penalize summaries that contained redundancies and excess information.

**Coherence** - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby “the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from a sentence to a coherent body of information about a topic.”

**Consistency** - the factual alignment between the summary and the summarized source. A factually consistent summary contains only statements that are entailed by the source document. Annotators were also asked to penalize summaries that contained hallucinated facts.

**Fluency** - the quality of the summary in terms of grammar, spelling, punctuation, word choice, and sentence structure.

You should output a single character to indicate which summary you think is better. ‘A’ stands for Summary A and ‘B’ stands for Summary B. If you think both summaries are equally good, output ‘E’.

Article:  
{article}

Summary A:  
{summary\_a}

Summary B:  
{summary\_b}

Your Choice (only a single character, you are allowed to think both summaries are equal and output ‘E’):

**GPT-4 Evaluation Prompt for HH-RLHF**

For the following query to a chatbot assistant, which response is more helpful?

First provide a one-sentence comparison of the two responses and explain which you feel is more helpful. Second, on a new line, state only ‘A’ or ‘B’ to indicate which response is more helpful. If they are equally good or bad, state ‘E’. Your response should use the json format, with “comparison” and “choice” as keys.

Query: dialogue

Response A: {resp\_a}

Response B: {resp\_b}

Your Judgment:

Table 12: Win rates on TL;DR using the Gemma-based models, varying the policy backbone and the embedding spaces used to form the cost matrix  $C$ .

Method	Backbone	Embedding	Win rate (vs. SFT-2B)	Win rate (vs. RKL-2B)
RKL-regularized PPO	Gemma-2B	-	0.848	-
	Gemma-7B	-	<b>0.948</b>	0.668
<b>Wasserstein-regularized PPO</b>	Gemma-2B	Gemma-2B	0.924	0.608
		Gemma-7B	0.908	0.556
	Gemma-7B	Gemma-2B	0.944	0.684
		Gemma-7B	<b>0.948</b>	<b>0.712</b>

Table 13: Sensitivity analysis of the truncation hyperparameter  $k_2$  on TL;DR with Gemma-2B. *Time* is the wall-clock time for the penalty computation.

$k_2$	Time (hours/1k steps)	Win rate (vs. SFT)	Win rate (vs. RKL)
64	0.08	0.864	0.528
128	0.12	0.924	0.608
256	0.19	0.916	0.584

## D ADDITIONAL EXPERIMENTAL RESULTS

### D.1 ANALYSIS OF EMBEDDING SPACE

The semantic cost matrix  $C$  is constructed from token embeddings and plays a central role in WPR. Since the cost must be computed over the full vocabulary of the policy model, the embedding space must be aligned with its tokenizer. Therefore, only models sharing the same tokenizer can be used directly. Using embeddings from a model with a different tokenizer would require building a cross-token alignment, a promising but nontrivial direction for future work.

To study the effect of embedding quality, we conduct an experiment using Gemma-2B and Gemma-7B, which share the same tokenizer. For each model, we extract the frozen token embeddings after SFT and used them to construct the cost matrix  $C$ . We then independently varied: (1) the policy backbone (Gemma-2B or Gemma-7B), and the embedding source used to form  $C$  (Gemma-2B SFT or Gemma-7B SFT).

As shown in Table 12, across all configurations, WPR consistently outperforms RKL-regularized PPO when using the same policy backbone, indicating that WPR provides benefits regardless of the specific embedding model used. As expected, the policy backbone size has the largest effect on performance, with Gemma-7B outperforming Gemma-2B.

Interestingly, constructing  $C$  using Gemma-7B embeddings for a Gemma-2B backbone does not yield performance improvements over using the 2B embeddings. We conjecture that this is because each policy is naturally grounded in the token geometry encoded by its own SFT embedding space. Thus, the embedding space of the same model backbone is most compatible with the policy’s internal representation.

### D.2 ADDITIONAL SENSITIVITY ANALYSIS

**Truncation hyperparameter  $k_2$**  To evaluate how the computational overhead and performance scale with the truncation hyperparameter  $k_2$ , we vary  $k_2 \in \{64, 128, 256\}$  and measure both the penalty computation time and the resulting win rates. The results are summarized in Table 13. Increasing  $k_2$  from 64 to 128 improves performance, with a moderate increase in penalty computation time, especially small compared to the overall training time of approximately 4.5 hours per 1,000 steps. Increasing  $k_2$  further from 128 to 256 yields minimal performance gains, and in fact slightly decreases performance. This suggests that  $k_2 = 128$  already captures most of the probability mass of the token distribution and provides an accurate approximation. Accordingly, we adopt  $k_2 = 128$  for all experiments in the paper.

Table 14: Sensitivity analysis of the entropy regularization parameter  $\lambda$  on TL;DR with Gemma-2B.

$\lambda$	Win rate (vs. SFT)	Win rate (vs. RKL)
50	0.900	0.564
100	0.924	0.608
200	0.916	0.612

Table 15: Detailed breakdown of the wall-clock time per 1,000 training steps. The time required to compute the regularization penalty differs across methods, whereas the generation and training steps are independent of the regularization method and therefore reported using unified timings.

	Time (hours/1k steps)	
	RKL	WPR
Generation		0.769
Penalty computation	0.005	0.117
Backpropagation		3.707
Total	4.481	4.593

**Entropy regularization hyperparameter  $\lambda$**  The hyperparameter  $\lambda$  controls the level of entropic smoothing in the Sinkhorn distance, determining the balance between the semantic fidelity and the smoothness of the transport plan. Smaller values of  $\lambda$  place greater weight on entropy, producing overly soft transport plans and diminishing the influence of semantic structure. Conversely, excessively large values of  $\lambda$  cause the kernel  $\mathbf{K} = \exp(-\lambda\mathbf{C})$  to collapse toward zero, creating numerical oscillations during the Sinkhorn rescaling steps.

To the best of our knowledge, this work is the first to apply an entropy-regularized Wasserstein penalty in the token space of LLMs during RL fine-tuning. Therefore, we initially selected  $\lambda$  empirically and found that  $\lambda = 100$  provided the stable and consistent performance across tasks. We use this value in all experiments reported in the paper. We additionally provide a sensitivity analysis with  $\lambda \in \{50, 100, 200\}$ . As shown in Table 14, WPR consistently outperforms RKL-based regularization for all tested values (with win rate is greater than 0.5). As expected, smaller values reduce the influence of semantic structure and lead to a performance drop, consistent with the interpretation above.

### D.3 COMPUTATIONAL RESOURCES

**Wall-clock Time** Table 15 reports a detailed breakdown of the wall-clock time per 1,000 PPO training steps for RKL and WPR. The measurement is decomposed into (1) generation, (2) penalty computation, and (3) backpropagation. For this report, we use 4 A100 GPUs using the Gemma-2B policy model with the TL;DR dataset, employing 8 batches per GPU and 8 gradient accumulation steps.

WPR requires additional computation during the penalty step because it incorporates semantic structure across tokens, whereas the KL penalty incurs nearly zero overhead. However, the added cost is minor relative to the forward and backward passes of a billion-parameter LLM. Note that the generation and backpropagation stages remain identical across regularization methods, and their runtimes are influenced far more by the generated response length, and therefore we report unified timings for these stages.

**GPU Memory Usage** The primary memory overhead of WPR arises from the cost matrix  $\mathbf{C}$ . Since  $\mathbf{C}$  is fixed across all training steps, it is computed once before training and reused throughout PPO optimization. Its memory footprint is independent of the model size.

As described in Section 4.2, we apply truncation methods that allow  $\mathbf{C}$  to be stored as a sparse matrix, requiring approximately 12.8GB of GPU memory in our configuration. This overhead is reflected in the peak GPU memory measurement in Table 16. During training, additional memory



Table 16: Peak GPU memory usage (GB) for RKL and Wasserstein regularization, measured on a single A100 GPU with a batch size of 8.

	GPU usage (GB)
RKL	64.05
<b>Wasserstein</b>	<b>78.98</b>

usage depends primarily on the batch size. When needed, memory can be reduced by increasing the gradient accumulation steps, at the cost of longer runtime.

## E GENERATED EXAMPLES

We present qualitative comparisons of response generation outputs. Table 17 shows an example from the TL;DR dataset, where we compare responses generated with reverse KL regularization and with our proposed Wasserstein Policy Regularization. Similarly, Table 18 provides an example from the HH-RLHF dataset under the same comparison. These examples illustrate that WPR produces responses that are better aligned with the underlying prompt compared to RKL.

## F LLM USAGE

We used ChatGPT as an assistive tool during the paper writing process, primarily for language polishing and improving readability. Additionally, we employed GPT-4o API in the experimental evaluation as an automatic judge to compare model responses and compute win rates, following established evaluation protocols.

Table 17: Comparison of response generation outputs from the prompt on the TL;DR dataset with RKL and WPR.

<b>Prompt</b>	<p>POST</p> <p>Subreddit: r/pettyrevenge</p> <p>I was driving home from work in San Antonio when I noticed this jerk in a sports car tailgating, honking, flashing lights, and in general being a major pest. We pulled up to a red light and he gunned it right through the intersection, somehow managing to not cause a major accident. I fumed for a bit but figured that was the end of it.</p> <p>Ten minutes later I pulled into my apartment complex gate (it had one of those bars that raises and lowers after you put in a code) and there he was, idling next to it. I don't know if he was just visiting so didn't have the right entry numbers or what, but he was just waiting for someone to come in so he could slip by on their pass.</p> <p>I pulled up to the gate, put my code in, and saw him inches from my bumper waiting to race in behind me. The gate raised up, I went forward, he followed intently, and I stopped with my rear bumper just barely in front of where the gate would fall.</p> <p>He fumed in his front seat while I waved, the gate lowered blocking him out, and I pulled away happy.</p> <p>TL;DR:</p>
<b>Responses</b>	
<b>RKL (A)</b>	26 year old male tailgating and harassing other motorists in San Antonio, TX causing minor fender bender when attempting to sneak past gate with illegal entry pass. Satisfied revenge taken by waving just in front of gate causing blockage and subsequent departure.
<b>WPR (ours) (B)</b>	jerk in sports car tailgating me, gunned it through intersection somehow avoiding major accident. Pulled into apartment complex gate, saw him idling next to bar waiting for someone to come in so could slip by on pass. Gate raised up blocking him out and I pulled away happy.
<b>GPT-4 Judgment</b>	B

Table 18: Comparison of response generation outputs from the prompt on the HH-RLHF dataset with RKL and WPR.

<b>Prompt</b>	Human: I need some assistance writing. See I was adopted years ago, and I am trying to meet my paternal mother. Can you write a letter asking if it is possible for us to meet?
	Assistant:
<b>Responses</b>	
<b>RKL (A)</b>	Yes, absolutely! Writing a letter to your biological mother asking for the opportunity to meet is a great way to initiate the process of connecting with your birth family. Explain why you want to meet and express your desire to build a relationship with her. Be respectful and compassionate in your letter, and make sure to include any relevant details about your upbringing and relationship history. Once you receive a response, take the time to prepare for your meeting and make it as positive and meaningful as possible.
<b>WPR (ours) (B)</b>	<p>Here is an example letter for requesting a meeting with your biological mother:</p> <p>Dear Mother,</p> <p>I would like to request a meeting to reconnect with you after many years of separation. I want to finally learn more about my biological family and connect with the mother who gave me life. Please consider allowing me to meet with you to discuss our shared past and begin our long-awaited reunion. Thank you.</p>
<b>GPT-4 Judgment</b>	Response B is more helpful because it provides a concrete example of a letter, which directly addresses the user’s request.