PREFERENCE-AWARE REWARD REDISTRIBUTION FOR TOKEN-LEVEL RLHF

Anonymous authors
Paper under double-blind review

ABSTRACT

Outcome reward models (ORMs) provide robust preference signals for reinforcement learning from human feedback (RLHF), yet their scalar response-level rewards leave the contribution of individual token decisions unspecified, creating a granularity mismatch with token-level policy optimization. We observe that the ORM itself provides a natural source of weak supervision for resolving the resulting ambiguity in mapping response-level rewards to individual tokens: the outcome reward is predicted from its terminal hidden representation, while intermediate token representations exhibit different degrees of alignment with this terminal state. We exploit this representational relation as a proxy for token-level preference relevance and propose Single-Source Entropic Optimal Transport (SEOT), a training-free framework that formulates outcome-to-token reward redistribution as a globally normalized allocation problem. SEOT models the terminal hidden state as a Dirac source and response-token hidden states as target supports, with their representational similarity defining preference-aware transportation costs. The resulting single-source entropic formulation admits a closed-form solution and requires neither additional reward model training nor iterative OT optimization, while exactly preserving the original outcome reward. Experiments on mathematical reasoning and general-domain tasks demonstrate consistent improvements over existing reward redistribution methods. Further analysis shows that SEOT better emphasizes preference-critical tokens while avoiding overly concentrated reward allocation, providing more effective token-level guidance for policy optimization.

1 INTRODUCTION

Reinforcement learning from human feedback (RLHF) has become a key paradigm for aligning Large Language Models (LLMs) with human preferences (Christiano et al., 2017; Glaese et al., 2022; Ouyang et al., 2022), where a preference reward model is trained and then frozen to guide policy optimization, most commonly with Proximal Policy Optimization (PPO) due to its relative stability (Schulman et al., 2017). Substantial progress in outcome reward models (ORMs), which assign a single scalar reward to completed responses, has been made through higher-quality preference data and more robust reward modeling to improve preference reward fidelity, stabilize policy optimization, and mitigate reward hacking (Gao et al., 2022; Liu et al., 2024; Li et al., 2025c; Duan et al., 2026; Liu et al., 2026; Yang et al., 2024; Coste et al., 2024; Li et al., 2026). However, regardless of how accurately an ORM captures response preferences, it still provides only a sparse and delayed response-level reward, whereas an autoregressive policy makes a sequence of token-level decisions (Chen et al., 2025; Chan et al., 2024; Yang et al., 2023; Li et al., 2025a; Cao et al., 2025). Consequently, a fundamental challenge is the granularity mismatch between token-level policy optimization and response-level terminal reward feedback, as sparse terminal rewards must be assigned across many token-level actions, often resulting in diluted or misallocated token-level reward.

A primary source of this difficulty is *Reward Redistribution Ambiguity* (Chan et al., 2024; Li et al., 2025a; Cao et al., 2025): a scalar outcome reward indicates the overall preference of a response but does not specify the contribution of individual token decisions. Standard PPO addresses a related but distinct problem through generalized advantage estimation (GAE), which propagates reward information temporally for advantage estimation rather than decomposing the terminal ORM reward into semantically meaningful token-level rewards (Schulman et al., 2018). Recent approaches therefore seek finer-grained learning signals through either token-level reward modeling (Yang et al.,

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

2023; Yoon et al., 2024; Chen et al., 2025; Yang et al., 2025) or reward redistribution (Chan et al., 2024; Zhou et al., 2024; Li et al., 2025a; Cao et al., 2025). The former learns additional token-level reward functions from response-level preference supervision, whereas the latter directly redistributes the reward produced by an existing ORM, thereby preserving its preference discrimination capability without additional reward model training. Existing training-free approaches such as ABC (Chan et al., 2024) estimate token importance from the ORM's attention patterns. However, such attribution signals do not explicitly characterize how individual token representations relate to the preference evidence underlying the final outcome reward. This leaves a central question for reward redistribution: how can we derive preference-aware token-level reward allocation directly from an existing ORM without explicit token-level supervision?

Our key observation is that the ORM itself provides a natural source of weak supervision for resolving this ambiguity. An ORM produces its outcome reward from the terminal hidden representation, which aggregates the preference evidence encoded throughout the response, while intermediate response token representations capture different degrees of alignment with this terminal preference representation. We therefore use their representational proximity as a training-free proxy for token-level preference relevance. This observation turns reward redistribution into a global allocation problem in which the outcome reward associated with a single terminal preference representation is distributed across response tokens according to their preference relevance. To this end, we propose **Single-Source Entropic Optimal Transport (SEOT)** based on the minimum-cost allocation principle of OT (Peyré & Cuturi, 2020; Benamou et al., 2014; Chizat et al., 2017; Courty et al., 2016). SEOT models the terminal hidden state as a Dirac source and response-token hidden states as target supports, with their representational similarity defining preference-aware transportation costs. Entropy regularization yields a globally normalized allocation that avoids degenerate concentration, while the single-source structure admits a closed-form solution without iterative OT optimization. The resulting token weights redistribute the terminal reward across response tokens while exactly preserving the original outcome reward, enabling plug-and-play token-level reward assignment for policy optimization without additional reward model training.

Our contributions are summarized as follows: (1) We develop a new reward redistribution method that can be built on existing robust ORMs to provide reliable token-level rewards for policy optimization without costly additional reward model training. (2) We formulate the redistribution of response-level rewards into reliable token-level rewards as a Single-Source Entropic Optimal Transport (SEOT) problem, where a transport cost derived from hidden-state similarity identifies preference-critical tokens and yields a closed-form solution without iterative optimization. (3) Extensive experiments on mathematical reasoning and general-domain tasks demonstrate consistent improvements over strong redistribution baselines.

2 Related Work

Token-Level Reward Assignment Recent advances in ORMs have produced increasingly robust reward models that mitigate reward hacking and preference bias (Duan et al., 2026; Li et al., 2025c; 2026), yet their response-level feedback still obscures individual token contributions, motivating fine-grained token reward for policy optimization. Existing approaches can be broadly divided into token-level reward modeling and terminal reward redistribution. Token-level reward modeling directly learns dense reward or value signals from response-level preferences, where representative methods such as PGG and QRM infer token-level reward by aggregating token-wise reward predictions or learning discriminative token-level value functions (Yang et al., 2023; Chen et al., 2025). Another line of work retains a pretrained ORM and redistributes its terminal reward without additional reward model training. ABC estimates token importance from reward model attention (Chan et al., 2024), RED derives token-level reward by redistributing holistic feedback through intermediate reward signals (Li et al., 2025a), and SCAR estimates marginal token contributions using Shapley values (Cao et al., 2025). While these methods provide convenient post-hoc reward redistribution, their effectiveness largely depends on the underlying attribution mechanism, which can produce inaccurate token importance or require costly repeated evaluations to estimate individual contributions (Chan et al., 2024; Cao et al., 2025). Our work instead revisits reward redistribution as an optimal transport problem for global probability-mass allocation, where the terminal hidden state is modeled as a single Dirac source distribution and the response token hidden states form a discrete target distribution.

We then derive normalized token-level rewards through a closed-form solution, requiring neither additional reward model training nor iterative contribution estimation.

Optimal Transport Optimal transport (OT) provides a principled framework for matching and reallocating probability mass across distributions under a minimum-cost objective (Peyré & Cuturi, 2020; Benamou et al., 2014; Chizat et al., 2017; Courty et al., 2016), and has been widely applied to generative models, computer vision, text analysis, and Long-tailed distribution (Genevay et al., 2017; Nguyen et al., 2021; Huynh et al., 2020; Gao et al., 2023; Guo et al., 2022). More recently, several studies have begun to integrate the advantages of OT into preference modeling to improve the robustness of outcome reward models (Li et al., 2025c; Pan et al., 2026) and into reinforcement learning algorithms for preference optimization (Li et al., 2025b). In contrast, our contribution is to use preference evidence encoded in hidden-state representations to define transport costs and reformulate the conventional transport between two discrete distributions as a single-source one-to-many allocation process, thereby deriving preference-aware token-level rewards while preserving the original preference capability of the pretrained ORM.

3 PRELIMINARIES

Language Generation as Sequential Decision-Making. We formulate autoregressive language generation as a token-level Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ denotes the textual state space, $\mathcal{A}$ the token action space, P the state transition, r the reward function, and $\gamma \in [0, 1)$ the discount factor. For an autoregressive language model π_ϕ, given an instruction $\boldsymbol{x}$ and a response $\boldsymbol{y} = (y_1, \ldots, y_T)$, the state $\boldsymbol{s}_t = (\boldsymbol{x}, \boldsymbol{y}_{<t})$ consists of the instruction and all previously generated response tokens, from which the next action $a_t = y_t$ is sampled according to $\pi_\phi(\cdot \mid \boldsymbol{s}_t)$. The sampled token extends the current prefix to $\boldsymbol{s}_{t+1} = (\boldsymbol{x}, \boldsymbol{y}_{\le t})$, and a complete response induces the trajectory $\boldsymbol{\tau} = (\boldsymbol{s}_1, a_1, \ldots, \boldsymbol{s}_T, a_T, \boldsymbol{s}_{T+1})$.

Outcome Reward Modeling and Reward Redistribution. A reward model (RM) learns a scalar proxy for human preference from pairwise preference data $\mathcal{D} = \{(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2)\}$, where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ denote the preferred and rejected responses to prompt $\boldsymbol{x}$, respectively. Let $\boldsymbol{\tau}_1 = (\boldsymbol{x}, \boldsymbol{y}_1)$ and $\boldsymbol{\tau}_2 = (\boldsymbol{x}, \boldsymbol{y}_2)$ denote the corresponding trajectories. A standard outcome reward model (ORM), parameterized by θ, models the preference probability under the Bradley–Terry (BT) objective (Bradley & Terry, 1952) as $p(\boldsymbol{\tau}_1 \succ \boldsymbol{\tau}_2 \mid \boldsymbol{x}) = \sigma(R_\theta(\boldsymbol{\tau}_1) - R_\theta(\boldsymbol{\tau}_2))$, where $R_\theta(\boldsymbol{\tau}) \in \mathbb{R}$ denotes the scalar response-level reward. Given a prompt $\boldsymbol{x} = (x_1, \ldots, x_m)$ and a response $\boldsymbol{y} = (y_1, \ldots, y_T)$, the ORM backbone produces hidden-state representations $(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_{m+T})$, where $\boldsymbol{h}_i \in \mathbb{R}^d$, and predicts the response-level reward from the terminal representation as $R_\theta(\boldsymbol{\tau}) = \boldsymbol{W}_r^\top \boldsymbol{h}_{m+T}$, where $\boldsymbol{W}_r \in \mathbb{R}^d$ denotes the reward-head parameters. Although recent ORMs have substantially improved robustness and generalization (Duan et al., 2026; Yang et al., 2024; Liu et al., 2026; Coste et al., 2024; Li et al., 2025c; 2026), the ORM still provides only a single scalar reward for the complete response, leaving the contribution of individual token decisions unspecified. We formulate token-level reward redistribution as a trajectory-dependent operator $\mathcal{F}$ that transforms the response-level reward into token-level rewards as $\boldsymbol{r}(\boldsymbol{\tau}) = \big(r_1(\boldsymbol{\tau}), \ldots, r_T(\boldsymbol{\tau})\big) \in \mathbb{R}^T$, where $r_t(\boldsymbol{\tau})$ denotes the reward assigned to token decision $(\boldsymbol{s}_t, a_t)$. For brevity, we write r_t when the dependence on $\boldsymbol{\tau}$ is clear from context. Unlike approaches that learn an additional token-level reward function, reward redistribution derives $\boldsymbol{r}(\boldsymbol{\tau})$ directly from the pretrained ORM without explicit token-level supervision (Chan et al., 2024; Li et al., 2025a; Cao et al., 2025). The central problem is therefore to design $\mathcal{F}$ such that the response-level reward is allocated across token decisions according to their preference relevance.

Optimal Transport. Optimal transport (OT) seeks a minimum-cost coupling between two probability distributions (Peyré & Cuturi, 2020; Benamou et al., 2014; Chizat et al., 2017; Courty et al., 2016). Consider two discrete distributions $P = \sum_{i=1}^{K} a_i \delta_{\boldsymbol{u}_i}$ and $Q = \sum_{j=1}^{L} b_j \delta_{\boldsymbol{v}_j}$, where $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ denote the source and target supports, and $\boldsymbol{a} = (a_1, \ldots, a_K)$ and $\boldsymbol{b} = (b_1, \ldots, b_L)$ denote their corresponding probability masses. Given a cost matrix $\boldsymbol{C} \in \mathbb{R}^{K \times L}$ with $C_{ij} = c(\boldsymbol{u}_i, \boldsymbol{v}_j)$, OT finds a transport plan $\boldsymbol{\Pi} \in \mathbb{R}_+^{K \times L}$ by solving

$$\boldsymbol{\Pi}^* = \arg\min_{\boldsymbol{\Pi} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{\Pi}, \boldsymbol{C} \rangle, \qquad U(\boldsymbol{a}, \boldsymbol{b}) = \left\{ \boldsymbol{\Pi} \ge 0 : \boldsymbol{\Pi} \mathbf{1}_L = \boldsymbol{a}, \boldsymbol{\Pi}^\top \mathbf{1}_K = \boldsymbol{b} \right\}. \tag{1}$$

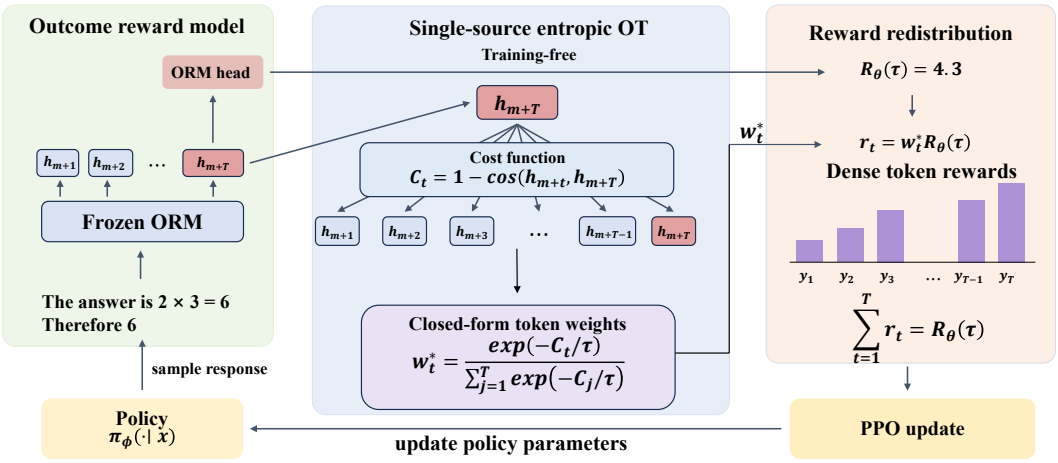


Figure 1: Overview of Single-Source Entropic Optimal Transport (SEOT): training-free redistribution of outcome rewards into preference-aware token-level rewards based on hidden-state similarity.

Here, Π_{ij} denotes the probability mass transported from $\boldsymbol{u}_i$ to $\boldsymbol{v}_j$, while $U(\boldsymbol{a}, \boldsymbol{b})$ enforces the source and target marginal constraints. Directly applying classical OT to reward redistribution is nontrivial because this setting does not provide a predefined pair of source and target probability distributions. More specifically, the ORM provides only a single response-level reward, while the token-level allocation required for redistribution is itself unknown. This motivates the single-source formulation introduced in the next section.

4 Single-Source Entropic Optimal Transport

In this section, we introduce Single-Source Entropic Optimal Transport (SEOT), a training-free framework for redistributing the scalar reward produced by a pretrained outcome reward model (ORM) into preference-aware token-level rewards. The overall framework is illustrated in Fig. 1.

4.1 Overall Method

We instantiate the reward redistribution operator $\mathcal{F}$ defined in Section 3 through **Single-Source Entropic Optimal Transport (SEOT)**. Our key observation is that the pretrained ORM itself provides a natural source of weak supervision for resolving reward redistribution ambiguity. Specifically, the outcome reward $R_\theta(\boldsymbol{\tau})$ is predicted from the terminal hidden-state $\boldsymbol{h}_{m+T}$, while the intermediate response-token hidden-states $\{\boldsymbol{h}_{m+t}\}_{t=1}^{T}$ encode different degrees of alignment with this terminal representation. We therefore use the representational relation between each response-token hidden-state and the terminal hidden-state as a proxy for its relevance to the final preference judgment. This differs from attention-based redistribution methods, which estimate token importance from internal attention patterns. Instead, SEOT anchors reward redistribution to the terminal representation directly used by the ORM to produce the outcome reward, and treats token-level reward allocation as a globally normalized allocation problem rather than independent token attribution. Based on this observation, we cast outcome-to-token reward redistribution as a single-source allocation problem. As illustrated in Fig. 1, SEOT uses hidden-state similarity to define preference-aware transportation costs, such that tokens more closely aligned with the terminal representation receive lower transport costs and larger normalized weights. SEOT realizes this allocation through an entropy-regularized transport formulation, which jointly determines the token weights while preserving the original outcome reward. The resulting weights redistribute the response-level preference signal into token-level rewards for subsequent policy optimization. We next introduce the specific transport formulation and derive its analytical solution.

4.2 SINGLE-SOURCE ENTROPIC OPTIMAL TRANSPORT

We instantiate the single-source allocation introduced in Section 4.1 by constructing a transport problem between the terminal hidden-state and response-token hidden-states. Since the ORM predicts the response-level reward from the terminal hidden-state $\boldsymbol{h}_{m+T}$, we model it as a Dirac source distribution with unit mass,

$$P = \delta_{\boldsymbol{h}_{m+T}}, \qquad \boldsymbol{a} = [1]. \tag{2}$$

Here, $\delta_{\boldsymbol{h}_{m+T}}$ denotes a Dirac distribution concentrated on the terminal hidden-state $\boldsymbol{h}_{m+T}$, such that the source distribution contains a single support with unit probability mass. The target distribution $Q_{\boldsymbol{w}}$ is supported on the hidden-state representations of the response tokens,

$$Q_{\boldsymbol{w}} = \sum_{t=1}^{T} w_t \delta_{\boldsymbol{h}_{m+t}}, \boldsymbol{w} \in \Delta_T, \tag{3}$$

where $\Delta_T = \left\{ \boldsymbol{w} \in \mathbb{R}_+^T : \sum_{t=1}^T w_t = 1 \right\}$. Here, $\boldsymbol{w} = (w_1, \ldots, w_T)$ denotes the unknown probability masses over the response-token hidden-state supports and is inferred by SEOT. Unlike classical OT, where both marginal distributions are prescribed, the target masses are not predefined in our setting. Since the source distribution contains only a single support with unit mass, each target mass is directly given by the corresponding transport mass, i.e., $w_t = \Pi_{1t}$, and subsequently serves as the token weight.

Preference-aware Transport Cost. The transport cost is a key component of the OT objective and can be flexibly specified according to the underlying problem. Based on our observation that the ORM hidden representations encode preference-relevant differences, we construct the transport cost from the representational relation between the terminal hidden-state and response-token hidden-states. Thus, response-token representations that are more closely aligned with the terminal representation should incur lower transportation costs and consequently receive larger target masses. Specifically, we use cosine distance to quantify the cost to the t-th target support as

$$C_t = 1 - \cos\left(\boldsymbol{h}_{m+t}, \boldsymbol{h}_{m+T}\right) = 1 - \frac{\boldsymbol{h}_{m+t}^\top \boldsymbol{h}_{m+T}}{\|\boldsymbol{h}_{m+t}\|_2 \|\boldsymbol{h}_{m+T}\|_2}. \tag{4}$$

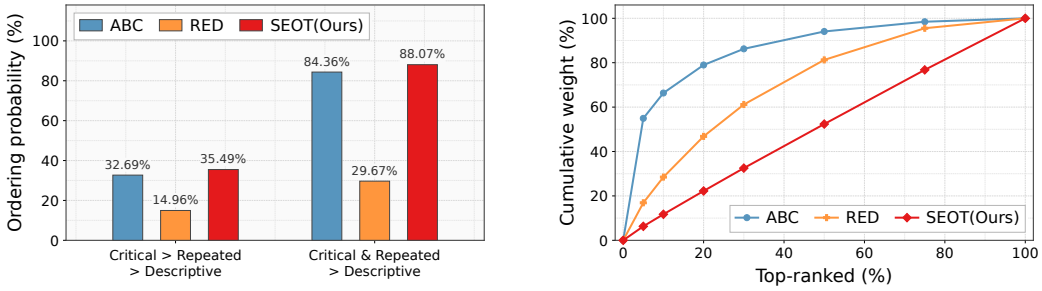
Entropic Single-Source Allocation. Although the preference-aware transport cost distinguishes response tokens according to their relevance to the terminal preference representation, directly minimizing the linear transport objective would concentrate all probability mass on the minimum-cost token. Such a degenerate allocation overlooks the preference information distributed across the response and provides supervision only to a small subset of tokens. We therefore introduce entropy regularization (Cuturi, 2013) to balance preference-aware allocation with broader token coverage, leading naturally to the following single-source entropic OT objective:

$$\boldsymbol{\Pi}^* = \arg\min_{\boldsymbol{\Pi} \in U(\boldsymbol{a})} \langle \boldsymbol{\Pi}, \boldsymbol{C} \rangle + \tau \sum_{t=1}^{T} \Pi_{1t} \log \Pi_{1t}, \quad U(\boldsymbol{a}) = \{\boldsymbol{\Pi} \geq 0 : \boldsymbol{\Pi} \mathbf{1}_T = \boldsymbol{a}\}, \quad \boldsymbol{a} = [1]. \tag{5}$$

where $\tau > 0$ controls the trade-off between minimizing the preference-aware transportation cost and distributing the mass across response tokens. Under the single-source formulation, the transport plan is fully characterized by the target masses, with $w_t = \Pi_{1t}$. Therefore, SEOT reduces to an entropy-regularized allocation over the T-dimensional probability simplex and admits the closed-form solution

$$w_t^* = \frac{\exp(-C_t/\tau)}{\sum_{j=1}^{T} \exp(-C_j/\tau)}, \qquad t = 1, \ldots, T. \tag{6}$$

Interestingly, the optimal allocation in Eq. 6 admits a softmax form over representation similarities. Specifically, this normalization is not introduced heuristically, but arises directly from the entropy-regularized single-source OT objective together with the probability-simplex constraint, which jointly preserve the total transport mass and control the concentration of the allocation. The complete derivation is provided in Appendix A. Consequently, the optimal target masses can be obtained directly without iterative OT optimization and subsequently serve as the token weights for reward redistribution.



(a) Weight effectiveness.

(b) Weight concentration.

Figure 2: Analysis of token-level reward weight on GSM8K. (a) Ordering rates between critical, repeated, and descriptive tokens. (b) Cumulative weight assigned to top-ranked tokens.

4.3 Reward Redistribution and Policy Optimization

In this subsection, we detail how the redistributed token-level rewards produced by SEOT are incorporated into PPO for policy optimization.

Training-Free Reward Redistribution. SEOT requires no additional reward model training and directly operates on a frozen ORM. For each sampled trajectory $\boldsymbol{\tau}$, a single forward pass through the ORM provides both the response-level reward $R_\theta(\boldsymbol{\tau})$ and the response-token hidden-states $\{\boldsymbol{h}_{m+t}\}_{t=1}^{T}$. These representations determine the preference-aware transportation costs and, through the closed-form SEOT solution, yield the normalized token weights $\boldsymbol{w}^*$ without iterative optimization. Since the ORM score reflects its original preference judgment, SEOT preserves its magnitude and changes only how the reward is distributed across token decisions, with $r_t(\boldsymbol{\tau}) = w_t^* R_\theta(\boldsymbol{\tau})$ and $\sum_{t=1}^{T} r_t(\boldsymbol{\tau}) = R_\theta(\boldsymbol{\tau})$. Therefore, SEOT derives token-level rewards directly from the original ORM without learning an additional token-level reward function or modifying the original preference signal.

Policy Optimization. Since each $r_t(\boldsymbol{\tau})$ is determined after the complete response is observed, it incorporates information from the future continuation through the terminal representation. For a fixed rollout policy π_ϕ, we characterize this trajectory-dependent signal by its conditional expectation $Z_{\pi_\phi}^{\mathrm{SEOT}}(\boldsymbol{s}_t, a_t) = \mathbb{E}_{\boldsymbol{\tau}_{>t}\sim\pi_\phi}\left[r_t(\boldsymbol{\tau}) \mid \boldsymbol{s}_t, a_t\right]$, for which the sampled $r_t(\boldsymbol{\tau})$ provides a Monte Carlo estimate. Accordingly, following QRM (Chen et al., 2025), we directly construct the token-level advantage as $\widehat{A}_t^{\mathrm{SEOT}} = r_t(\boldsymbol{\tau}) - V_\psi(\boldsymbol{s}_t)$, where V_ψ denotes the learned value function, and optimize the policy with the standard clipped PPO objective without GAE. The complete training procedures for SEOT and GAE-free PPO are provided in Appendix B.

5 Experiments

5.1 Experimental Setup

Training and Evaluation Datasets. We evaluate SEOT across two domains, including mathematical reasoning and cross-domain generalization. For mathematical reasoning, we construct preference pair datasets from GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) following QRM (Chen et al., 2025) for Reward Modeling. We then optimize the policy on GSM8K and MATH using PPO (Schulman et al., 2017) and evaluate it on their corresponding in-domain (ID) test sets. For the generalization ability, we directly adopt the strong pretrained Skywork-Reward-Llama-3.1-8B reward model (Liu et al., 2024) trained on the Skywork-Reward-Preference-80K-v0.2 (SP) training set (Liu et al., 2025), and perform PPO training on Alpaca-GPT4-En (Peng et al., 2023) with evaluation on a diverse benchmark including AlpacaEval 2.0 (Dubois et al., 2025), as well as out-of-distribution (OOD) evaluation on MATH-500 (Lightman et al., 2023) and RACE (Lai et al., 2017). Detailed descriptions of all datasets are provided in Appendix C.1.

Table 1: Mathematical reasoning performance of policy models on GSM8K and MATH, measured by Pass@N (%). Training indicates whether the reward model requires additional training. Best performances are shown in **bold**, while suboptimal ones are underlined. Superscript deltas indicate changes relative to PPO+ORM for each model.

Methods	Training	GSM8K			MATH			Avg
		Pass@1	Pass@8	Pass@16	Pass@1	Pass@8	Pass@16	Pass@1
Llama3.2-3B-instruct								
SFT	—	74.63	94.63	96.97	26.46	61.81	70.12	50.55
PPO+ORM	✓	75.58	**94.96**	**97.04**	25.44	58.64	67.10	50.51
PPO+DPO-RM	✓	74.27$^{\downarrow 1.31}$	94.60$^{\downarrow 0.36}$	96.82$^{\downarrow 0.22}$	26.22$^{\uparrow 0.78}$	62.18$^{\uparrow 3.54}$	71.10$^{\uparrow 4.00}$	50.25$^{\downarrow 0.26}$
PPO+QRM	✓	74.00$^{\downarrow 1.58}$	93.17$^{\downarrow 1.79}$	95.60$^{\downarrow 1.44}$	26.25$^{\uparrow 0.81}$	61.90$^{\uparrow 3.26}$	70.68$^{\uparrow 3.58}$	50.13$^{\downarrow 0.38}$
PPO+ABC-ORM	✗	74.88$^{\downarrow 0.70}$	94.33$^{\downarrow 0.63}$	96.36$^{\downarrow 0.68}$	26.07$^{\uparrow 0.63}$	61.80$^{\uparrow 3.16}$	70.42$^{\uparrow 3.32}$	50.48$^{\downarrow 0.03}$
PPO+RED-ORM	✗	73.98$^{\downarrow 1.60}$	93.67$^{\downarrow 1.29}$	95.75$^{\downarrow 1.29}$	25.96$^{\uparrow 0.52}$	61.64$^{\uparrow 3.00}$	70.42$^{\uparrow 3.32}$	49.97$^{\downarrow 0.54}$
PPO+SEOT-ORM (ours)	✗	**78.77**$^{\uparrow 3.19}$	94.51$^{\downarrow 0.45}$	96.59$^{\downarrow 0.45}$	**28.61**$^{\uparrow 3.17}$	**63.52**$^{\uparrow 4.88}$	**71.86**$^{\uparrow 4.76}$	**53.69**$^{\uparrow 3.18}$
Qwen3.5-2B								
SFT	—	73.70	94.18	96.29	44.93	66.91	71.66	59.32
PPO+ORM	✓	74.43	94.02	**96.36**	44.78	66.81	71.68	59.61
PPO+QRM	✓	**81.94**$^{\uparrow 7.51}$	**94.57**$^{\uparrow 0.55}$	**96.36**$^{\uparrow 0.00}$	**60.62**$^{\uparrow 15.84}$	**78.77**$^{\uparrow 11.96}$	**82.48**$^{\uparrow 10.80}$	**71.28**$^{\uparrow 11.67}$
PPO+ABC-ORM	✗	73.60$^{\downarrow 0.83}$	93.65$^{\downarrow 0.37}$	95.83$^{\downarrow 0.53}$	44.90$^{\uparrow 0.12}$	66.70$^{\downarrow 0.11}$	71.42$^{\downarrow 0.26}$	59.25$^{\downarrow 0.36}$
PPO+RED-ORM	✗	73.75$^{\downarrow 0.68}$	94.14$^{\uparrow 0.12}$	96.29$^{\downarrow 0.07}$	45.46$^{\uparrow 0.68}$	67.16$^{\uparrow 0.35}$	71.96$^{\uparrow 0.28}$	59.61$^{\uparrow 0.00}$
PPO+SEOT-ORM (ours)	✗	76.19$^{\uparrow 1.76}$	94.03$^{\uparrow 0.01}$	96.13$^{\downarrow 0.23}$	51.50$^{\uparrow 6.72}$	73.24$^{\uparrow 6.43}$	77.48$^{\uparrow 5.80}$	63.85$^{\uparrow 4.24}$

Base Models and Training Details. We retrain BT-ORMs on the preference pairs solely to provide outcome rewards, while SEOT itself requires no additional training. For mathematical reasoning, we initialize the reward models from Llama-3-8B-Instruct (Meta, 2024) and Qwen3.5-2B [1] and fine-tune them with Low-Rank Adaptation (LoRA) (Hu et al., 2021). We use Llama-3.2-3B-Instruct [2] and Qwen3.5-2B as the corresponding policy models, with tokenizer-compatible reward and policy models to ensure consistent token-level alignment. For general-domain instruction following, we use the pretrained Skywork-Reward-Llama-3.1-8B (Liu et al., 2024) as the reward model and Llama-3.2-3B-Instruct as the policy model. All policy models are evaluated using EvalScope[3]. Training objectives, hyperparameters, and implementation details are provided in Appendix C.2.

Baselines. As detailed in Appendix C.3, we compare with ORM and three types of reward distribution methods: (1) ORM (Bradley & Terry, 1952); (2) training-based methods, including QRM (Chen et al., 2025) and DPO-RM (Rafailov et al., 2024a;b); and (3) redistribution-based methods, including ABC (Chan et al., 2024) and RED (Li et al., 2025a).

5.2 Performance Evaluation

Mathematical reasoning ability. To evaluate mathematical reasoning ability, we fine-tune Llama-3.2-3B-Instruct and Qwen3.5-2B with PPO on GSM8K and MATH and evaluate the resulting policies using Pass@N, the probability of obtaining at least one correct answer from N sampled responses. As shown in Table 1, SEOT achieves the highest Pass@1 on Llama-3.2-3B-Instruct, reaching 78.77% and 28.61% on GSM8K and MATH, respectively. It consistently outperforms the redistribution baselines ABC and RED and also exceeds the training-based QRM by 4.77 and 2.36 points. A similar improvement over ORM, ABC, and RED is observed on Qwen3.5-2B, where SEOT achieves substantial gains on the more challenging MATH benchmark, improving Pass@1 by 6.72 points over ORM and by 6.60/6.04 points over ABC/RED. Although QRM performs better on Qwen3.5-2B, SEOT consistently outperforms the training-free redistribution baselines across both backbones without requiring additional reward model training.

Performance analysis of Pass@k. Our observations closely align with recent findings on RLVR (Yue et al., 2025), which show that RL training mainly improves Pass@1 by increasing the probability of sampling rewarded trajectories, while its benefit can diminish at larger k. A similar trend is observed for SEOT. On GSM8K, where SFT already achieves high Pass@8 and Pass@16, SEOT mainly improves Pass@1, and even the training-based QRM exhibits the same saturation effect. In contrast, MATH starts from substantially lower Pass@8 and Pass@16, leaving more room for

[1] https://huggingface.co/Qwen/Qwen3.5-2B
[2] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct
[3] https://github.com/modelscope/evalscope

Table 2: Performance (%) of Llama-3.2-3B-Instruct policies trained on Alpaca-GPT4-En with Skywork-Reward-Llama-3.1-8B as the reward model. Best performances are shown in **bold**, while suboptimal ones are underlined. Superscript deltas indicate changes relative to PPO+ORM.

Methods	AlpacaEval 2.0	MATH-500	RACE	Avg
SFT	12.80	38.80	57.74	36.45
PPO+ORM	15.16	44.40	58.80	39.45
PPO+QRM	14.53$^{\downarrow 0.63}$	43.80$^{\downarrow 0.60}$	64.86$^{\uparrow 6.06}$	41.06$^{\uparrow 1.61}$
PPO+ABC-ORM	13.66$^{\downarrow 1.50}$	41.40$^{\downarrow 3.00}$	57.54$^{\downarrow 1.26}$	37.53$^{\downarrow 1.92}$
PPO+RED-ORM	13.66$^{\downarrow 1.50}$	41.60$^{\downarrow 2.80}$	63.91$^{\uparrow 5.11}$	39.72$^{\uparrow 0.27}$
PPO+SEOT-ORM (ours)	**15.90**$^{\uparrow 0.74}$	**48.00**$^{\uparrow 3.60}$	**66.20**$^{\uparrow 7.40}$	**43.37**$^{\uparrow 3.92}$

improvement, and SEOT further improves these metrics from 61.81/70.12 to 63.52/71.86 on Llama-3.2-3B-Instruct and from 66.91/71.66 to 73.24/77.48 on Qwen3.5-2B. Although SEOT relies on preference-based token-level rewards rather than binary verifiable rewards, both settings share a similar optimization mechanism when the ORM correctly distinguishes preferable responses. By emphasizing critical reasoning tokens, SEOT provides finer-grained guidance that primarily benefits Pass@1 after high-k saturation but can still improve Pass@8 and Pass@16 on challenging benchmarks.

General-domain ability. We evaluate general-domain instruction following and OOD generalization on AlpacaEval 2.0, MATH-500, and RACE. As shown in Table 2, SEOT achieves 15.90, 48.00, and 66.20 on the three benchmarks, respectively, with an average score of 43.37, corresponding to improvements of 0.74, 3.60, and 7.40 points over PPO+ORM. These gains are larger than those achieved by both the training-free redistribution methods ABC and RED and the training-based QRM, with SEOT attaining the strongest OOD performance on MATH-500 and RACE. The comparison with ABC further indicates that preserving the original ORM reward magnitude alone is insufficient, as effective redistribution also requires effective token-level allocation. In contrast, RED modifies the response-level reward magnitude during redistribution, which may affect the transfer of the pretrained ORM's original preference signal across domains. Overall, SEOT combines preservation of the original ORM reward with more effective preference-aware allocation, leading to stronger and more consistent generalization.

GSM8K performance under GAE. Our main experiments adopt the GAE-free formulation to more directly examine the effect of token-level reward assignment on policy optimization without additional temporal reward propagation from GAE. We also further evaluate ABC, RED, and SEOT under the same conventional GAE-based PPO setting. As shown in Table 3, SEOT achieves a Pass@1 of 77.07%, outperforming ABC and RED by 2.37 and 1.46 points, respectively. Although SEOT remains effective under GAE and consistently outperforms ABC and RED, its Pass@1 decreases from 78.77% to 77.07%, suggesting that temporal credit propagation may dilute the fine-grained reward guidance provided by SEOT.

5.3 Result Analysis

Reward redistribution analysis. We further analyze whether SEOT produces preference-aware token weights without excessive concentration. Unlike ABC, which estimates token importance from attention patterns, SEOT anchors redistribution to the terminal preference representation and performs globally normalized allocation. Using the same GSM8K responses and a fixed reward model, we compare SEOT with ABC and RED based on token ordering consistency and weight concentration. Since ground-truth token-level rewards are unavailable, we use the following semantic categories as task-specific proxies. We define critical tokens as the first occurrence of the final result, repeated tokens as later restatements of that result, and descriptive tokens as ordinary descriptions or intermediate computations. For RED, we normalize its token rewards within each response for comparison, and average weights when a numerical result spans multiple tokens. As shown in Fig. 2a, the ordering Critical $>$ Repeated $>$ Descriptive holds in 35.49% of cases for SEOT, compared with 32.69% for ABC and 14.96% for RED. Meanwhile, both critical and repeated tokens receive larger weights than descriptive tokens in 88.07% of cases, versus 84.36% and 29.67%, respectively. Fig. 2b further shows that SEOT distributes weight more broadly, with the top 10% tokens accounting for only 11.72% of the total weight, compared with 66.34% for ABC, while RED exhibits an intermediate level of concen-

Table 3: GSM8K Pass@1 performance under GAE-based PPO.

Methods	GSM8K Pass@1
PPO+ABC-ORM	74.70
PPO+RED-ORM	75.61 $^{\uparrow 0.91}$
PPO+SEOT-ORM (ours)	**77.07** $^{\uparrow 2.37}$

Table 4: Preservation of pretrained ORM preference discrimination on RewardBench.

Method	Chat	Chat Hard	Safety	Reasoning	Overall
ORM	**94.7**	**88.4**	**92.7**	**96.7**	**93.1**
SEOT	**94.7**	**88.4**	**92.7**	**96.7**	**93.1**
QRM	91.6 $^{\downarrow 3.1}$	73.9 $^{\downarrow 14.5}$	90.8 $^{\downarrow 1.9}$	91.9 $^{\downarrow 4.8}$	87.1 $^{\downarrow 6.0}$

tration. These results suggest that SEOT retains preference-sensitive token ordering while avoiding the highly concentrated allocation produced by attention-based attribution, yielding a smoother global reward allocation. A complete example with token-wise weights is provided in Appendix D.

Stability of PPO Optimization. SEOT is designed to emphasize critical tokens while retaining appropriate weight for descriptive tokens, a balance that should support stable PPO optimization. We assess this expectation by tracking Pass@1 throughout PPO training. As shown in Fig. 3, SEOT maintains stable performance, whereas QRM, DPO-RM, ABC, and RED degrade as training proceeds. This behavior is consistent with SEOT's more balanced reward allocation observed above, suggesting that more reliable emphasis on critical tokens may contribute to stable policy updates throughout training. Overall, although ABC and RED also provide token-level rewards, their relatively more concentrated reward distributions may contribute to greater performance degradation under GAE-free optimization.

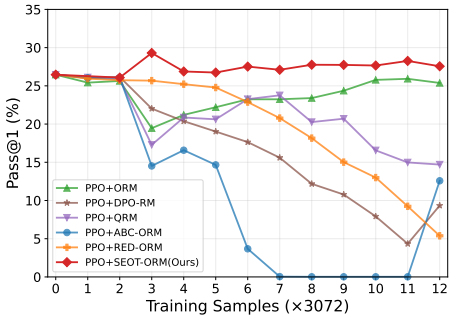


Figure 3: Training stability of PPO for Llama-3.2-3B-Instruct on MATH.

Preserving outcome-level preference discrimination. Modern open-source reward models have achieved strong preference discrimination performance (Duan et al., 2026; Li et al., 2025c; 2026; Liu et al., 2024), making it important to determine whether such capabilities can be retained when deriving token-level rewards. To examine this, we evaluate the strong pretrained Skywork-Reward-Llama-3.1-8B on RewardBench (Lambert et al., 2024), which measures preference discrimination across chat, chat-hard, safety, and reasoning scenarios. Notably, QRM requires further training to remodel the pretrained ORM into a token-level reward model, whereas SEOT keeps the ORM unchanged and directly redistributes its response-level reward. Consequently, as shown in Table 4, QRM reduces the original RewardBench accuracy from 93.1% to 87.1%, whereas SEOT preserves the original 93.1% performance by construction. Combined with the downstream policy gains in Table 2, this result indicates that SEOT improves policy optimization not by modifying the underlying preference model, but by more effectively reallocating its existing preference signal across token decisions.

6 Conclusion

In this work, we observe that similarity differences between the terminal hidden-state and response-token hidden-states provide implicit weak supervision for reward redistribution, motivating an optimal transport formulation and our Single-Source Entropic Optimal Transport (SEOT) method. SEOT treats the terminal hidden-state as a Dirac source and response-token hidden-states as target supports, defining preference-aware transport costs from representational similarity to obtain a closed-form, training-free reward allocation that preserves the original outcome reward. Across mathematical reasoning and general-domain tasks, SEOT yields more effective token-level rewards than existing redistribution methods, improving policy performance and optimization stability.

Limitations. The current implementation of SEOT assumes tokenizer-compatible reward and policy models to directly align ORM representations with policy token decisions. In addition, SEOT uses representation similarity to the terminal ORM state as a proxy for token-level preference relevance, and its effectiveness therefore depends on the quality of the preference information encoded in the ORM representations. Extending SEOT to heterogeneous reward-policy architectures and more general preference representations remains an interesting direction for future work.

AI USE STATEMENT

Large language models (LLMs) were used only to assist with the writing of this manuscript, including translating the authors' intended meaning into appropriate academic English, correcting grammatical and typographical errors, improving sentence structure and clarity, and addressing formatting and typesetting issues. The scientific content of this work, including its motivation, methodology, experimental design, analysis, and conclusions, was developed by the authors. No LLM-assisted output was incorporated directly without human review. All such output was carefully checked, revised, and verified by the authors before inclusion in the manuscript.

REFERENCES

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems, 2014. URL https://arxiv.org/abs/1412.5154.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952. URL https://api.semanticscholar.org/CorpusID:125209808.

Meng Cao, Shuyuan Zhang, Xiao-Wen Chang, and Doina Precup. Scar: Shapley credit assignment for more efficient rlhf, 2025. URL https://arxiv.org/abs/2505.20417.

Alex J. Chan, Hao Sun, Samuel Holt, and Mihaela van der Schaar. Dense reward for free in reinforcement learning from human feedback, 2024. URL https://arxiv.org/abs/2402.00782.

Hongzhan Chen, Tao Yang, Shiping Gao, Ruijun Chen, Xiaojun Quan, Hongtao Tian, and Ting Yao. Discriminative policy optimization for token-level reward models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=aq3YxKPZBk.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems, 2017. URL https://arxiv.org/abs/1607.05816.

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. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

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

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dcjtMYkpXx.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation, 2016. URL https://arxiv.org/abs/1507.00504.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013. URL https://arxiv.org/abs/1306.0895.

Zhibin Duan, Guowei Rong, Zhuo Li, Bo Chen, Mingyuan Zhou, and Dandan Guo. Mitigating reward hacking in RLHF via bayesian non-negative reward modeling. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=DfhMMHXDuu.

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

Jintong Gao, He Zhao, Zhuo Li, and Dandan Guo. Enhancing minority classes by mixing: An adaptative optimal transport approach for long-tailed classification. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 60329–60348. Curran Associates, Inc., 2023. doi: 10.52202/075280-2636. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/bdabb5d4262bcfb6a1d529d690a6c82b-Paper-Conference.pdf.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization, 2022. URL https://arxiv.org/abs/2210.10760.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences, 2017. URL https://arxiv.org/abs/1706.00292.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements, 2022. URL https://arxiv.org/abs/2209.14375.

Dandan Guo, Long Tian, He Zhao, Mingyuan Zhou, and Hongyuan Zha. Adaptive distribution calibration for few-shot learning with hierarchical optimal transport, 2022. URL https://arxiv.org/abs/2210.04144.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-Rank Adaptation of Large Language Models, 2021. URL https://arxiv.org/abs/2106.09685.

Viet Huynh, He Zhao, and Dinh Phung. Otlda: A geometry-aware optimal transport approach for topic modeling. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18573–18582. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d800149d2f947ad4d64f34668f8b20f6-Paper.pdf.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Rewardbench: Evaluating reward models for language modeling, 2024. URL https://arxiv.org/abs/2403.13787.

Jiahui Li, Lin Li, Tai wei Chang, Kun Kuang, Long Chen, Jun Zhou, and Cheng Yang. Red: Unleashing token-level rewards from holistic feedback via reward redistribution, 2025a. URL https://arxiv.org/abs/2411.08302.

Meng Li, Guangda Huzhang, Haibo Zhang, Xiting Wang, and Anxiang Zeng. Optimal transport-based token weighting scheme for enhanced preference optimization, 2025b. URL https://arxiv.org/abs/2505.18720.

Zhuo Li, Yuege Feng, Dandan Guo, Jinpeng Hu, Anningzhe Gao, and Xiang Wan. Aplot: Robust reward modeling via adaptive preference learning with optimal transport, 2025c. URL https://arxiv.org/abs/2510.10963.

Zhuo Li, Pengyu Cheng, Zhechao Yu, FeifeiTong, Anningzhe Gao, Tsung-Hui Chang, Xiang Wan, erchao.zec, xiaoxi jiang, and guanjunjiang. Eliminating inductive bias in reward models with information-theoretic guidance. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=57YfUhcYXd.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. URL https://arxiv.org/abs/2305.20050.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms, 2024. URL https://arxiv.org/abs/2410.18451.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-Reward-V2: Scaling Preference Data Curation via Human-AI Synergy, 2025. URL https://arxiv.org/abs/2507.01352.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and Yang Liu. Skywork-reward-v2: Scaling preference data curation via human-AI synergy. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ofgxkMLqic.

Meta. The llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783.

Tuan Nguyen, Trung Le, He Zhao, Quan Hung Tran, Truyen Nguyen, and Dinh Phung. Most: multi-source domain adaptation via optimal transport for student-teacher learning. In Cassio de Campos and Marloes H. Maathuis (eds.), *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 225–235. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/nguyen21a.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Licheng Pan, Haochen Yang, Haoxuan Li, Yunsheng Lu, Yongqi Tong, Yinuo Wang, Shijian Wang, Zhixuan Chu, Lei Shen, Yuan Lu, and Hao Wang. Optimal transport for llm reward modeling from noisy preference, 2026. URL https://arxiv.org/abs/2605.06036.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4, 2023. URL https://arxiv.org/abs/2304.03277.

Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020. URL https://arxiv.org/abs/1803.00567.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a q-function, 2024a. URL https://arxiv.org/abs/2404.12358.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024b. URL https://arxiv.org/abs/2305.18290.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2018. URL https://arxiv.org/abs/1506.02438.

Kailai Yang, Zhiwei Liu, Qianqian Xie, Jimin Huang, Erxue Min, and Sophia Ananiadou. Selective preference optimization via token-level reward function estimation, 2025. URL https://arxiv.org/abs/2408.13518.

Rui Yang, Ruomeng Ding, Yong Lin, Huan Zhang, and Tong Zhang. Regularizing hidden states enables learning generalizable reward model for llms, 2024. URL https://arxiv.org/abs/2406.10216.

Shentao Yang, Shujian Zhang, Congying Xia, Yihao Feng, Caiming Xiong, and Mingyuan Zhou. Preference-grounded token-level guidance for language model fine-tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=6SRE9GZ9s6.

Eunseop Yoon, Hee Suk Yoon, SooHwan Eom, Gunsoo Han, Daniel Nam, Daejin Jo, Kyoung-Woon On, Mark Hasegawa-Johnson, Sungwoong Kim, and Chang Yoo. TLCR: Token-level continuous reward for fine-grained reinforcement learning from human feedback. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14969–14981, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.889. URL https://aclanthology.org/2024.findings-acl.889/.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. URL https://arxiv.org/abs/2504.13837.

Wenxuan Zhou, Shujian Zhang, Lingxiao Zhao, and Tao Meng. T-reg: Preference optimization with token-level reward regularization, 2024. URL https://arxiv.org/abs/2412.02685.

A Derivation of the Analytical Solution of SEOT

Since the source distribution contains only one support with unit probability mass, the transport plan $\mathbf{\Pi} \in \mathbb{R}_+^{1\times T}$ contains only one row and can be written as $\{\Pi_{1t}\}_{t=1}^T = \{w_t\}_{t=1}^T$. Substituting it into the SEOT objective in Eq. 5 reduces the original transport problem to

$$\begin{aligned} &\min_{\boldsymbol{w}} \sum_{t=1}^{T} w_t C_t + \tau \sum_{t=1}^{T} w_t \log w_t, \\ &\text{s.t.} \quad \sum_{t=1}^{T} w_t = 1, \qquad w_t \geq 0, \quad t = 1, \ldots, T, \end{aligned} \tag{7}$$

where the first term minimizes the transportation cost while the entropy regularization prevents the probability mass from collapsing onto a single target support. To derive the analytical solution, we introduce a Lagrange multiplier λ for the normalization constraint and obtain

$$\mathcal{L}(\boldsymbol{w}, \lambda) = \sum_{t=1}^{T} w_t C_t + \tau \sum_{t=1}^{T} w_t \log w_t + \lambda \left(\sum_{t=1}^{T} w_t - 1 \right).$$

Taking the derivative with respect to w_t gives

$$\frac{\partial \mathcal{L}}{\partial w_t} = C_t + \tau(\log w_t + 1) + \lambda = 0,$$

which yields

$$w_t = \exp\left(-\frac{C_t}{\tau}\right) \exp\left(-1 - \frac{\lambda}{\tau}\right).$$

Using the normalization condition $\sum_{t=1}^{T} w_t = 1$, we obtain

$$\exp\left(-1 - \frac{\lambda}{\tau}\right) = \frac{1}{\sum_{j=1}^{T} \exp\left(-\frac{C_j}{\tau}\right)}.$$

Therefore,

$$w_t^* = \Pi_{1t}^* = \frac{\exp\left(-C_t/\tau\right)}{\sum_{j=1}^{T} \exp\left(-C_j/\tau\right)}, \qquad t = 1, \ldots, T. \tag{8}$$

Finally, substituting the transportation cost $C_t = 1 - \cos(\boldsymbol{h}_{m+t}, \boldsymbol{h}_{m+T})$ gives

$$w_t^* = \frac{\exp\left(-\left[1 - \cos(\boldsymbol{h}_{m+t}, \boldsymbol{h}_{m+T})\right]/\tau\right)}{\sum_{j=1}^{T} \exp\left(-\left[1 - \cos(\boldsymbol{h}_{m+j}, \boldsymbol{h}_{m+T})\right]/\tau\right)}.$$

Since the common factor $\exp(-1/\tau)$ cancels during normalization, this simplifies to

$$w_t^* = \frac{\exp\left(\cos(\boldsymbol{h}_{m+t}, \boldsymbol{h}_{m+T})/\tau\right)}{\sum_{j=1}^{T} \exp\left(\cos(\boldsymbol{h}_{m+j}, \boldsymbol{h}_{m+T})/\tau\right)}, \qquad t = 1, \ldots, T. \tag{9}$$

Thus, the single-source SEOT problem admits an exact closed-form solution without requiring iterative OT optimization.

B SEOT and PPO Training Procedure

This section provides the complete training procedures for SEOT and GAE-free PPO. Algorithm 1 summarizes the SEOT reward redistribution process. Algorithm 2 further describes how the resulting token-level rewards are incorporated into GAE-free PPO for policy optimization.

Algorithm 1 Single-Source Entropic Optimal Transport (SEOT)

Input: Prompt $\boldsymbol{x}$ of length m, response $\boldsymbol{y} = (y_1, \ldots, y_T)$, frozen ORM parameterized by θ, entropy coefficient $\tau > 0$
Ensure: Token-level rewards $\{r_t\}_{t=1}^T$

$R_\theta(\boldsymbol{\tau}), \{\boldsymbol{h}_{m+t}\}_{t=1}^T \leftarrow \mathrm{Forward}(R_\theta, \boldsymbol{x}, \boldsymbol{y})$ ▷ Response-level reward and hidden-states
$C_t \leftarrow 1 - \cos(\boldsymbol{h}_{m+t}, \boldsymbol{h}_{m+T}), \quad t = 1, \ldots, T$ ▷ Preference-aware transport costs
$w_t^* \leftarrow \dfrac{\exp(-C_t/\tau)}{\sum_{j=1}^T \exp(-C_j/\tau)}, \quad t = 1, \ldots, T$ ▷ Closed-form SEOT solution
$r_t(\boldsymbol{\tau}) \leftarrow w_t^* R_\theta(\boldsymbol{\tau}), \quad t = 1, \ldots, T$ ▷ Reward redistribution
return $\{r_t(\boldsymbol{\tau})\}_{t=1}^T$ ▷ $\sum_{t=1}^T r_t(\boldsymbol{\tau}) = R_\theta(\boldsymbol{\tau})$

Algorithm 2 GAE-Free PPO with SEOT

Input: PPO dataset $\mathcal{D}_{\mathrm{PPO}}$, initial policy π_0, frozen ORM parameterized by θ, entropy coefficient τ, clipping parameter ϵ, KL coefficient β_{KL}, learning rates η_π, η_V

Initialize policy $\pi_\phi \leftarrow \pi_0$, reference policy $\pi_{\mathrm{ref}} \leftarrow \pi_0$, and value function V_ψ
repeat
  Set $\pi_{\mathrm{old}} \leftarrow \pi_\phi$ and sample a batch of responses $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ from $\mathcal{D}_{\mathrm{PPO}}$ using π_{old}
  for each response $\boldsymbol{y}_i$ **do**
    Compute token-level rewards $\{r_{i,t}\} \leftarrow \mathrm{SEOT}(\boldsymbol{x}_i, \boldsymbol{y}_i, R_\theta, \tau)$ using Algorithm 1
    Compute values $v_{i,t} = V_\psi(\boldsymbol{s}_{i,t})$
  end for
  Compute standard deviations σ_r, σ_v over all valid response tokens in the batch
  Scale rewards and values as $\bar{r}_{i,t} = \frac{r_{i,t} - \mathrm{mean}(r)}{\sigma_r}$ and $\bar{v}_{i,t} = \frac{v_{i,t} - \mathrm{mean}(v)}{\sigma_v}$
  for each response $\boldsymbol{y}_i$ **do**
    Compute reference-policy penalty $k_{i,t} = \left| \log \frac{\pi_{\mathrm{old}}(a_{i,t}|\boldsymbol{s}_{i,t})}{\pi_{\mathrm{ref}}(a_{i,t}|\boldsymbol{s}_{i,t})} \right|$
    Compute value target $G_{i,t} = \bar{r}_{i,t} - \beta_{\mathrm{KL}} k_{i,t}$
    Compute GAE-free advantage $\widehat{A}_{i,t}^{\mathrm{SEOT}} = G_{i,t} - \bar{v}_{i,t}$
  end for
  Compute probability ratio $\rho_{i,t}(\phi) = \frac{\pi_\phi(a_{i,t}|\boldsymbol{s}_{i,t})}{\pi_{\mathrm{old}}(a_{i,t}|\boldsymbol{s}_{i,t})}$
  Compute clipped surrogate loss $\mathcal{L}_\pi = -\,\mathrm{mean}\Big[\min\Big(\rho_{i,t}\widehat{A}_{i,t}^{\mathrm{SEOT}}, \mathrm{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)\widehat{A}_{i,t}^{\mathrm{SEOT}}\Big)\Big]$
  Compute value loss $\mathcal{L}_V = \mathrm{MSE}(V_\psi(\boldsymbol{s}_{i,t}), G_{i,t})$
  Update ϕ over actor minibatches using Adam with learning rate η_π to minimize $\mathcal{L}_\pi$
  Update ψ over critic minibatches using Adam with learning rate η_V to minimize $\mathcal{L}_V$
until the prescribed training batches have been processed

C Single Source Entropic Optimal Transport Training Details

C.1 Datasets

We now describe the datasets used in our experiments. For mathematical reasoning, GSM8K (Cobbe et al., 2021) evaluates grade-school mathematical reasoning and contains 7,473 training problems and 1,319 test problems, while MATH (Hendrycks et al., 2021) evaluates competition-level mathematical reasoning and contains 7,500 training problems and 5,000 test problems. Following (Chen et al., 2025), we further construct 45,364 and 94,941 preference pairs from their training sets, respectively. For cross-domain comprehension, Skywork-Reward-Preference-80K-v0.2 (SP) (Liu et al., 2025) contains 77,016 preference pairs covering diverse general-domain instructions and responses, while Alpaca-GPT4-En (Peng et al., 2023) contains 52,002 English instruction-following examples generated by GPT-4, from which we randomly shuffle and select 20,000 examples for our experiments. For policy evaluation in this setting, AlpacaEval 2.0 (Dubois et al., 2025) evaluates general-domain instruction following on 805 instructions, for which we report the standard Win Rate. MATH-500 (Lightman et al., 2023) contains a subset of 500 problems from MATH to evaluate mathematical reasoning generalization, while RACE (Lai et al., 2017) evaluates reading comprehension using English examination passages, for which we use the test split containing 4,934 questions.

Table 5: Hyperparameter settings for reward modeling and PPO training.

Hyperparameter	Reward Modeling		PPO Training	
	Llama-3-8B	Qwen3.5-2B	Llama-3.2-3B	Qwen3.5-2B
Base model	Llama-3-8B-Instruct	Qwen3.5-2B	Llama-3.2-3B-Instruct	Qwen3.5-2B
Training strategy	LoRA	LoRA	Full-parameter	Full-parameter
Max sequence length	1024	1024	1024/2048	1024
Batch size	2	2	10	5
Learning rate	1×10^{-5}	1×10^{-5}	5×10^{-6}	5×10^{-6}
Warmup ratio	–	–	0.05	0.05
Epoch	1	1	1	1
τ	1	0.5	1	0.5
LoRA rank r	128	128	–	–
LoRA α	128	128	–	–
Value LoRA rank r	–	–	128	128
Value LoRA α	–	–	128	128
BT scale β_{BT}	0.2	0.2	–	–
BT margin γ_{BT}	2.0	2.0	–	–

Skywork-Reward-Llama-3.1-8B is directly adopted as the pretrained reward model for the general-domain experiments and therefore requires no additional reward-model training.

Table 6: Tokenization statistics on identical response texts using different tokenizers.

Dataset	Llama-3.2-3B-Instruct	Qwen3.5-2B	Increase
Llama SEOT GSM8k responses	150.71	169.73	+12.62%
Qwen SEOT GSM8k responses	295.13	333.73	+13.08%

C.2 Training Hyperparameters

We train the reward and value models with LoRA while updating the policy model with full-parameter fine-tuning. The BT-ORMs are fine-tuned for one epoch using the margin Bradley–Terry objective adopted in QRM (Chen et al., 2025) as $\mathcal{L}_{\mathrm{BT}} = -\log \sigma\left(\beta_{\mathrm{BT}}[R_\theta(\boldsymbol{x}, \boldsymbol{y}_1) - R_\theta(\boldsymbol{x}, \boldsymbol{y}_2)] - \gamma_{\mathrm{BT}}\right)$, where we set $\beta_{\mathrm{BT}} = 0.2$ and $\gamma_{\mathrm{BT}} = 2.0$ to encourage greater score separation between chosen and rejected responses. QRM is further trained from the corresponding base model using the same objective.

Both reward modeling and PPO use a maximum context length of 1024 tokens on GSM8K and MATH, whereas QRM and policy training on Alpaca-GPT4-En use 2048 tokens. All remaining hyperparameters are summarized in Table 5. For SEOT, the entropy coefficient τ controls the sharpness of the token weight allocation. We set $\tau = 1$ for Llama-3.2-3B-Instruct and $\tau = 0.5$ for Qwen3.5-2B. The smaller value for Qwen helps maintain sufficient allocation contrast because its tokenizer produces longer token sequences for the same response, causing the normalized reward mass to be distributed across more token positions, as shown in Table 6.

For GAE-based experiments, following Q-RM (Chen et al., 2025), we set the GAE parameter λ to 0.8, the PPO discount factor γ to 0.9, and the KL coefficient to 0.01. During evaluation, we use the EvalScope framework with its default settings under the corresponding context-length limits, while the Qwen model is evaluated exclusively in the zero-shot setting because its finer-grained tokenizer produces longer token sequences and some prompts already approach the context limit before generation. We further report results on AlpacaEval 2.0 (Dubois et al., 2025), MATH-500 (Lightman et al., 2023), and RACE (Lai et al., 2017) to assess instruction-following, OOD mathematical reasoning, and reading comprehension abilities, respectively. For AlpacaEval 2.0, we use the DeepSeek-V4 Pro API as the LLM judge. All evaluated policy models are deployed locally with vLLM[4] and accessed through an API for evaluation.

[4] https://github.com/vllm-project/vllm

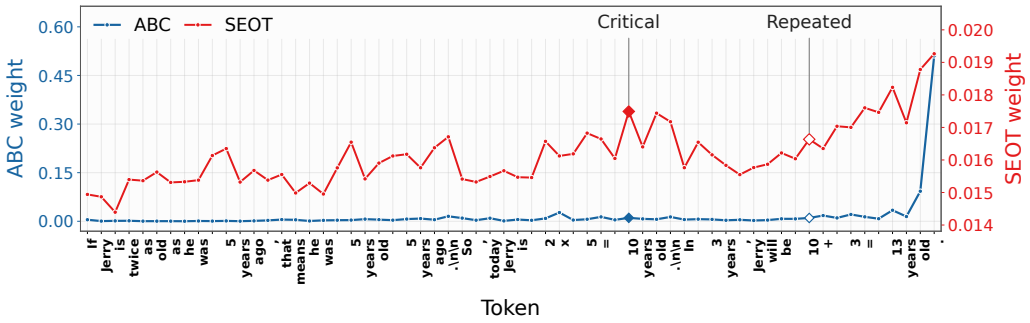


Figure 4: Token-level reward redistribution of SEOT on GSM8K.

C.3 Introduction to Baselines

We experimentally consider the following response-level, training-based token-level, and reward redistribution baselines:

1. **ORM** (Bradley & Terry, 1952), a classical ranking-based preference model that assigns a scalar reward to the entire response and uses GAE to estimate token-level advantages for policy optimization.
2. **QRM** (Chen et al., 2025), which decouples reward modeling from language generation and optimizes a discriminative policy to learn token-level Q-function rewards directly from response-level preference data without fine-grained annotations.
3. **DPO-RM** (Rafailov et al., 2024a;b), which derives token-level rewards from the implicit reward model induced by DPO as $r_{\mathrm{DPO}}(s_t, a_t) = \beta \log \frac{\pi(a_t|s_t)}{\pi_{\mathrm{ref}}(a_t|s_t)}$.
4. **ABC** (Chan et al., 2024), which uses the attention weights of the reward model to redistribute its response-level reward across generated tokens, providing token-level rewards without additional RM training.
5. **RED** (Li et al., 2025a), which applies the same ORM to different response prefixes and computes token-level rewards by taking the score difference $r_t = R_\theta(x, y_{\leq t}) - R_\theta(x, y_{\leq t-1})$.

D Example of SEOT's transportation plan

This subsection discusses token-level reward distributions produced by SEOT and representative method ABC on GSM8K, MATH, and RewardBench as shown in Fig. 4–6. Across all three examples, ABC concentrates most weight on a small number of tokens, often producing sharp peaks near the end of the response, whereas SEOT maintains broader coverage while assigning larger weights to preference-critical token. Moreover, the first occurrence of key reasoning or answer content tends to receive higher SEOT weight than its later repetitions, suggesting that SEOT captures contextual preference relevance rather than merely token position or lexical repetition. Overall, the transportation plan assigns appropriate weights to descriptive tokens while emphasizing both initial and repeated occurrences of preference-critical information, thereby providing balanced token-level guidance for stable and effective PPO optimization.

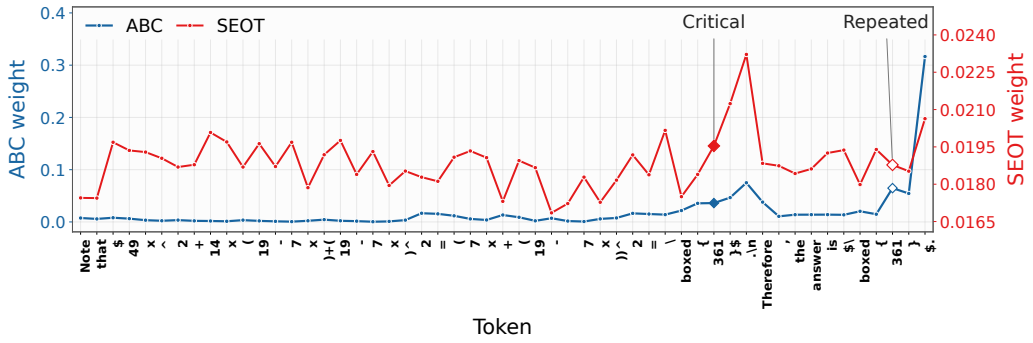


Figure 5: Token-level reward redistribution of SEOT on MATH.

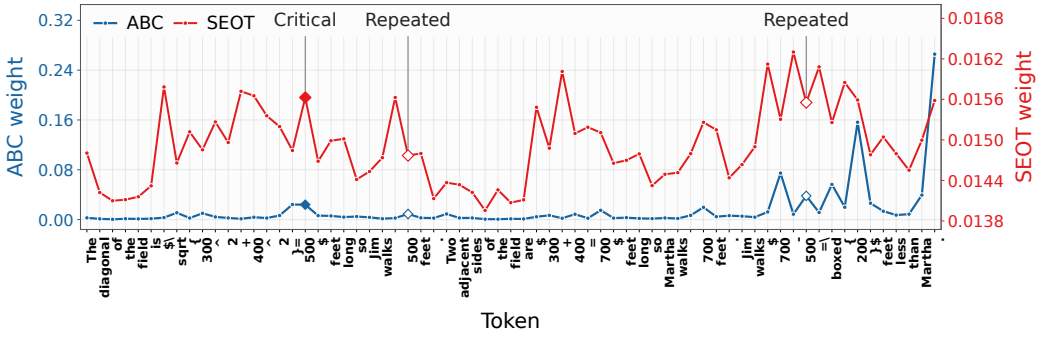


Figure 6: Token-level reward redistribution of SEOT on RewardBench.