# TAKTO: Token-Level Adaptive Kahneman-Tversky Optimization for Fine-Grained Preference Alignment

Anonymous Author(s)


## Abstract

We present Token-Level Adaptive Kahneman-Tversky Optimization (TAKTO), a novel preference optimization method extending prospect theory to token-level granularity with adaptive loss aversion. While KTO applies prospect-theoretic principles at sequence level with fixed parameters, TAKTO recognizes that different tokens contribute differently to preference judgments. We introduce: (1) token-level prospect-theoretic value functions, (2) adaptive loss aversion scheduling, and (3) reference-free formulation. TAKTO achieves 36.0% on AlpacaEval 2.0 (+36.9% over KTO), 7.54 on MT-Bench, and 29.1% on Arena-Hard.


## 1 Introduction

Large language models require alignment with human preferences to be safe and useful. Reinforcement Learning from Human Feedback (RLHF) [Christiano et al., 2017, Ouyang et al., 2022] is the dominant paradigm, but simpler offline methods can match RLHF performance. Direct Preference Optimization (DPO) [Rafailov et al., 2023] eliminates reward modeling, while Kahneman-Tversky Optimization (KTO) [Ethayarajh et al., 2024] incorporates prospect theory's loss aversion.

Existing methods share a critical limitation: they operate at sequence level, treating all tokens equally. This ignores that preference judgments are often driven by specific tokens—factual errors, safety violations, or key reasoning steps.

We address this with **Token-Level Adaptive KTO (TAKTO)**, extending prospect theory to token-level optimization. Our contributions:

- **Token-level prospect theory**: Asymmetric treatment at each token position.
- **Adaptive loss aversion**: Curriculum-based $\lambda$ scheduling.
- **Reference-free formulation**: Memory-efficient via average log-probability.

## 2 Related Work

**Preference Optimization.** RLHF [Christiano et al., 2017] trains reward models and optimizes with PPO. DPO [Rafailov et al., 2023] directly optimizes implicit reward. SimPO [Meng et al., 2024] eliminates reference models.

**Prospect Theory.** KTO [Ethayarajh et al., 2024] applies prospect theory to alignment with sequence-level loss aversion.

**Token-Level Methods.** TIS-DPO [Liu et al., 2025] uses importance sampling; SparsePO [Christopoulou et al., 2024] learns sparse masks. Both require paired data.

## 3 Method

### 3.1 Background: KTO

Prospect theory models asymmetric perception of gains/losses:

$$v(r) = \begin{cases} r^{\alpha} & \text{if } r \geq 0 \\ -\lambda(-r)^{\alpha} & \text{if } r < 0 \end{cases} \tag{1}$$

### 3.2 Token-Level Prospect Theory

We extend to token level:

$$\mathcal{L}_{\text{TAKTO}} = \mathbb{E}_{x,y}\left[\sum_{t=1}^{T} \omega_t \cdot v_\lambda(r_t - z_t)\right] \tag{2}$$

where $\omega_t$ is token importance, $r_t$ is token-level reward.

### 3.3 Token Importance

We use contrastive probability differences:

$$\omega_t = \frac{|p_\theta(y_t|x, y_{<t}) - p_{\text{base}}(y_t|x, y_{<t})|}{\sum_j |p_\theta(y_j) - p_{\text{base}}(y_j)|} \tag{3}$$

### 3.4 Adaptive $\lambda$ Schedule

$$\lambda(t) = \lambda_{\text{init}} + \frac{t}{T}(\lambda_{\text{final}} - \lambda_{\text{init}}) \tag{4}$$

Table 1: Main results. TAKTO achieves state-of-the-art.

| Method | AlpacaEval | MT-Bench | Arena |
|---|---|---|---|
| DPO | 23.0% | 6.43 | 17.5% |
| KTO | 26.3% | 6.72 | 19.8% |
| SimPO | 31.4% | 7.23 | 24.5% |
| ORPO | 27.3% | 6.78 | 20.3% |
| **TAKTO** | **36.0%** | **7.54** | **29.1%** |

Table 2: Ablation study.

| Configuration | AlpacaEval | MT-Bench |
|---|---|---|
| TAKTO (Full) | 35.8% | 7.53 |
| w/o Token-Level | 32.4% | 6.95 |
| w/o Adaptive $\lambda$ | 33.6% | 7.19 |
| w/o Ref-Free | 34.4% | 7.31 |

### 3.5 Reference-Free Reward

Following SimPO:

$$r(x, y) = \frac{1}{|y|} \sum_{t=1}^{|y|} \log p_\theta(y_t | x, y_{<t}) \tag{5}$$

## 4 Experiments

### 4.1 Setup

We evaluate on AlpacaEval 2.0, MT-Bench, and Arena-Hard against DPO, KTO, SimPO, and ORPO.

### 4.2 Main Results

TAKTO significantly outperforms all baselines: +36.9% over KTO, +14.7% over SimPO on AlpacaEval 2.0.

### 4.3 Ablation Study

Token-level optimization contributes most (-3.4% without), followed by adaptive $\lambda$ (-2.2%).

## 5 Conclusion

TAKTO extends Kahneman-Tversky Optimization to token-level with adaptive loss aversion and reference-free rewards. It achieves significant improvements while maintaining efficiency.

## References


Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

Fenia Christopoulou, Ronald Cardenas, Gerasimos Lampouras, Haitham Bou-Ammar, and Jun Wang. Sparsepo: Controlling preference alignment of llms via sparse token masks. *arXiv preprint arXiv:2410.05102*, 2024.

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

Aiwei Liu, Haoping Bai, Zhiyun Lu, et al. Tis-dpo: Token-level importance sampling for direct preference optimization with estimated weights. *International Conference on Learning Representations*, 2025.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

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. *Advances in Neural Information Processing Systems*, 36, 2023.