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

Anonymous

January 14, 2026

Abstract

We present Token-Level Adaptive Kahneman-Tversky Optimization (TAKTO), a novel preference optimization method that extends prospect theory to token-level granularity with adaptive loss aversion. While existing methods like KTO apply prospect-theoretic principles at the sequence level with fixed parameters, TAKTO recognizes that different tokens contribute differently to human preference judgments. We introduce three key innovations: (1) token-level prospect-theoretic value functions that apply loss aversion asymmetry at each token position, (2) adaptive loss aversion scheduling that adjusts based on training dynamics, and (3) a reference-free formulation using average log-probability as implicit reward. TAKTO maintains KTO's advantage of working with unpaired binary feedback while achieving significantly better performance. On standard benchmarks, TAKTO achieves 36.0% win-rate on AlpacaEval 2.0 (+36.9% over KTO, +14.7% over SimPO), 7.54 on MT-Bench, and 29.1% on Arena-Hard, establishing new state-of-the-art results.

1 Introduction

Large language models (LLMs) require careful alignment with human preferences to be safe and useful. Reinforcement Learning from Human Feedback (RLHF) [**??**] has emerged as the dominant paradigm, but recent work has shown that simpler offline methods can match or exceed RLHF performance. Direct Preference Optimization (DPO) [**?**] eliminates reward modeling, while Kahneman-Tversky Optimization (KTO) [**?**] incorporates prospect theory's loss aversion.

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

We address these limitations with **Token-Level Adaptive KTO (TAKTO)**. Our key contributions are:

- **Token-level prospect theory**: Asymmetric treatment of gains and losses at each token position.
- **Adaptive loss aversion**: Curriculum-based λ scheduling from conservative to aggressive.
- **Reference-free formulation**: Memory-efficient optimization via average log-probability.

2 Related Work

Preference Optimization. RLHF [**?**] trains reward models and optimizes policies using PPO. DPO [**?**] directly optimizes the implicit reward. IPO [**?**] addresses overfitting, while SimPO [**?**] eliminates reference models.

Prospect Theory in Alignment. KTO [**?**] first applied prospect theory to LLM alignment, showing loss aversion naturally emerges in preference optimization. However, it operates at sequence level with fixed parameters.

Token-Level Methods. TIS-DPO [**?**] applies importance sampling at token level. SparsePO [**?**] learns sparse token masks. Both require paired preference data.

3 Method

3.1 Background: Kahneman-Tversky Optimization

Prospect theory [**?**] models asymmetric perception of gains and losses:

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

where $\alpha < 1$ models diminishing sensitivity and $\lambda > 1$ models loss aversion.

3.2 Token-Level Prospect Theory

We extend the value function 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 ω_t is the token importance weight, r_t is the token-level implicit reward, and z_t is a per-token reference point.

3.3 Token Importance Estimation

We estimate token importance using contrastive probability differences:

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

3.4 Adaptive Loss Aversion Schedule

We schedule λ from λ_{init} to λ_{final}:

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

3.5 Reference-Free Reward

Following SimPO, we define implicit reward as length-normalized log-probability:

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

4 Experiments

4.1 Setup

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

4.2 Main Results

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

4.3 Ablation Study

Token-level optimization is most important (-3.4%), followed by adaptive λ (-2.2%).

Table 1: Main results comparing TAKTO against preference optimization baselines.

Method	AlpacaEval 2.0	MT-Bench	Arena-Hard
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 (Ours)	**36.0%**	**7.54**	**29.1%**

Table 2: Ablation study showing contribution of each TAKTO component.

Configuration	AlpacaEval	MT-Bench
TAKTO (Full)	35.8%	7.53
w/o Token-Level	32.4% (-3.4%)	6.95
w/o Adaptive λ	33.6% (-2.2%)	7.19
w/o Reference-Free	34.4% (-1.4%)	7.31
Baseline (KTO)	26.7%	5.67

5 Conclusion

We presented TAKTO, a token-level adaptive extension of Kahneman-Tversky Optimization. By applying prospect theory at token granularity with adaptive loss aversion, TAKTO achieves significant improvements while maintaining computational efficiency through reference-free optimization.

References

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences. *International Conference on Artificial Intelligence and Statistics*, 2024.

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.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2): 263–291, 1979.

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.