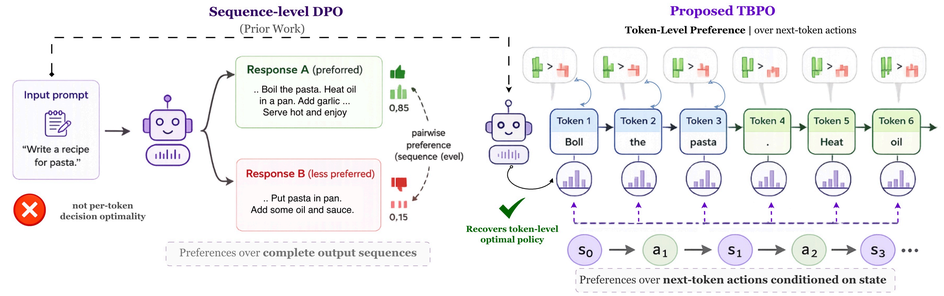
TokenRatio: Principled Token-Level Preference Optimization via Ratio Matching

Truong Nguyen[1*], Tien-Phat Nguyen[1*], Linh Ngo Van[1], Duy Minh Ho Nguyen[2,3,4], Khoa D. Doan[5], Trung Le[6]

1 Hanoi University of Science & Technology · 2 University of Stuttgart · 3 DFKI · 4 Max Planck Research School (IMPRS-IS) · 5 VinUniversity · 6 Monash University

* Equal contribution · Correspondence: linhnv@soict.hust.edu.vn, trunglm@monash.edu

Code & models
github.com/dinhtruongng/TBPO



Sequence-level DPO supervises **whole responses**; our **TBPO** models preferences over **next-token actions conditioned on the state** — recovering a token-level optimal policy from the same pairwise data.

At a glance

GOAL

Recover a **token-level optimal policy** using only standard **sequence-level pairwise comparisons** — no reward model, no on-policy RL.

CONTRIBUTIONS

- **TBPO** — a token-level preference framework that models preferences per token and learns from sequence-level pairs.
- **Theory** — minimizing the TBPO objective provably recovers a token-level optimal policy.
- **Results** — stronger alignment, better stability, and higher diversity than strong sequence- and token-level baselines.

60.9	70.5	+5pt
Mistral-7B Open-LLM avg	Llama-3-8B Open-LLM avg	ARC & GSM8K vs. baselines

01 MOTIVATION — The Problem

LLMs generate **one token at a time**, yet preference objectives score **whole sequences**.

THE MISMATCH

- DPO fits a **sequence-level** BT model over complete responses.
- Optimality should hold at **every state** — but supervision is sequence-level.
- Early errors **cascade** — per-token credit diluted across the whole sequence.

PRIOR TOKEN-LEVEL METHODS STILL AREN'T

- **TDPO** & **TIS-DPO** redistribute one sequence-level signal across timesteps.
- The underlying preference model stays sequence-level — **per-state optimality is only implicit**.

GENERATION AS A TOKEN-LEVEL MDP

$$s_t^w = [x, y_{<t}^w], \quad s_t^l = [x, y_{<t}^l]$$

$$p_{\text{data}}(a_t^w \succ a_t^l \mid s_t^w, s_t^l) = \sigma\big(S(s_t^w, a_t^w) - S(s_t^l, a_t^l)\big)$$

$$\max_{\pi_\theta} \mathbb{E}\big[A_{\pi_{\text{ref}}}([x, y_{<t}], z) - \beta\, D_{\text{KL}}\big(\pi_\theta(\cdot \mid s) \,\|\, \pi_{\text{ref}}(\cdot \mid s)\big)\big]$$

$$p_{\text{data}}(y_{\le t}^w \succ y_{\le t}^l \mid x) = \prod_{i=1}^{t} p_{\text{data}}(y_i^w \succ y_i^l \mid x, y_{<i}^w, y_{<i}^l)$$

OUR FIX

Lift the **Bradley–Terry model itself** to the token level — match a per-token preference ratio so optimality holds at **every state** from the same sequence-level pairs.

$$\max_{\pi_\theta} \mathbb{E}\big[A_{\pi_{\text{ref}}}([x, y_{<t}], z) - \beta\, D_{\text{KL}}\big(\pi_\theta(\cdot \mid s) \,\|\, \pi_{\text{ref}}(\cdot \mid s)\big)\big]$$

02 METHOD — TBPO

Token-level Bradley–Terry model

Lift each sequence label $y^w \succ y^l$ to a per-step comparison of next-token actions, each taken under **its own prefix**:

$$s_t^w = [x, y_{<t}^w], \quad s_t^l = [x, y_{<t}^l]$$

$$p_{\text{data}}(a_t^w \succ a_t^l \mid s_t^w, s_t^l) = \sigma\big(S(s_t^w, a_t^w) - S(s_t^l, a_t^l)\big)$$

TBPO-Q

score = state–action value $S(s,a) = Q_{\pi_{\text{ref}}}(s,a)$

Correction w_t is a **partition-function ratio**, fit by a lightweight baseline head (stop-grad on the LM).

$$w_t^{(Q)} = \frac{Z([x, y_{<t}^l]; \beta)}{Z([x, y_{<t}^w]; \beta)}$$

TBPO-A

score = advantage $S(s,a) = A_{\pi_{\text{ref}}}(s,a)$

Partition function cancels; w_t becomes a **KL-baseline difference**, estimated with a low-variance K3 estimator.

$$w_t^{(A)} = \exp\big(D_{\text{KL}}^l - D_{\text{KL}}^w\big)$$

Bregman ratio matching → tractable loss

$$R_\theta^t = \left(\frac{\pi_\theta(y_t^l \mid s^l)\, \pi_{\text{ref}}(y_t^w \mid s^w)}{\pi_{\text{ref}}(y_t^l \mid s^l)\, \pi_\theta(y_t^w \mid s^w)}\, w_t\right)^\beta$$

$$\mathcal{L}(\theta) = \mathbb{E}\, \frac{1}{T} \sum_{t=1}^{T} \Big[h'(R_\theta^t) R_\theta^t - h(R_\theta^t) - h'\big((R_\theta^t)^{-1}\big)\Big]$$

Scaled Basu's power divergence (SBA) generator — λ trades off confident vs. uncertain samples for **robustness to noisy preferences**:

$$h(R) = \frac{R^{1+\lambda} - R}{s\,\lambda(\lambda+1)}$$

OPTIMALITY (THEOREM)

Under sufficient capacity, the minimizer of the token-level ratio-matching objective **is the token-level optimal policy** of the KL-regularized objective.

$$\pi_{\theta^*} = \arg\min_{\pi_\theta} D_h(R_{\text{data}}, R_\theta)$$

03 EVALUATION — Experiments

60.95	70.50	+5 pts
Open-LLM avg, Mistral-7B best overall (DPO 59.30)	**Open-LLM avg**, Llama-3-8B best overall (TDPO 67.91)	**ARC & GSM8K** gains vs. token-level baselines

GENERAL CAPABILITIES · OPEN LLM LEADERBOARD

METHOD	HELLASW.	ARC	MMLU	TRUTHQA	WINOGR.	GSM8K	AVG
Mistral 7B v0.1							
SFT	80.7	55.5	58.4	43.7	76.7	18.5	**55.6**
DPO	82.4	59.4	59.2	42.3	77.7	34.9	**59.3**
TDPO	82.7	**61.5**▪	57.2	43.5	77.0	28.7	**58.4**
TIS-DPO	**83.1**▪	59.1	57.8	**45.9**▪	77.7	34.4	**59.7**
BPO	81.0	57.7	59.2	41.7	77.3	32.9	**58.3**
TBPO-Q	82.7	59.7	**60.8**▪	44.5	**78.6**▪	**39.3**▪	**60.9**▪
TBPO-A	82.7	59.5	60.5	44.4	78.2	39.0	**60.7**
Llama 3 8B							
SFT	79.5	51.5	61.2	46.5	75.7	68.7	**63.8**
DPO	81.2	56.1	61.7	48.5	75.8	74.2	**66.2**
TDPO	**83.3**▪	59.0	61.9	51.8	76.6	75.1	**67.9**
TIS-DPO	81.4	58.7	60.8	49.9	75.1	75.1	**66.8**
BPO	81.7	55.0	61.5	47.9	75.1	73.2	**65.7**
TBPO-Q	81.9	64.1	**64.9**▪	**53.4**▪	79.1	78.9	**70.4**
TBPO-A	82.0	**64.1**▪	64.8	53.3	**79.6**▪	**79.2**▪	**70.5**▪

Higher is better. ▪ best per column. TBPO leads the **overall average on both backbones**, with the largest gains on reasoning (GSM8K, ARC).

SETUP

Mistral-7B & Llama-3-8B, SFT-init. Judges: Llama-3-70B, DeepSeek-V3, Mixtral-8×22B.

WHY TBPO WINS

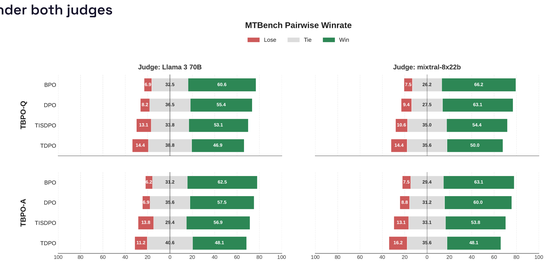
- Biggest gains on **reasoning** (ARC, GSM8K **+5 pts**).
- Wins at **shortest output length** — not from verbosity.
- Recovers **generation diversity** lost by other PO methods.

$$\pi_{\theta^*} = \arg\min_{\pi_\theta} D_h(R_{\text{data}}, R_\theta)$$

GENERATION QUALITY · HELPFULNESS · DIVERSITY
TBPO wins more, stays concise, and keeps outputs diverse

MT-BENCH · PAIRWISE

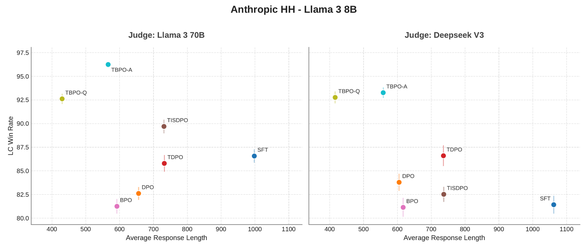
Higher win rates under both judges



TBPO-Q/A win the majority of **decisive** comparisons against every baseline, consistently across judges.

ANTHROPIC HH · LLAMA-3-8B

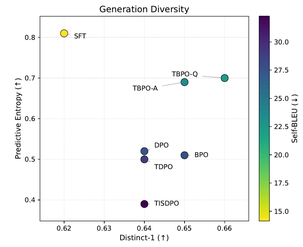
Best helpfulness — with shorter answers



TBPO sits top-left: **highest LC win rate at the shortest length**, so gains aren't from verbosity.

GENERATION DIVERSITY

Recovers diversity lost by PO



TBPO pushes the frontier up-right: **higher entropy & Distinct-1, lower Self-BLEU** than other PO methods.