TUR-DPO: Topology- and Uncertainty-Aware Direct Preference Optimization

Abdulhady Abas[1] Fatemeh Daneshfar[2] Seyedali Mirjalili[3,4] Mourad Oussalah[5]

[1]University of Kurdistan, Erbil [2]University of Kurdistan, Iran [3]Torrens University Australia
[4]Óbuda University, Hungary [5]University of Oulu, Finland

ICML 2026




Topics of Presentation

Presentation Topics

1. Background: RLHF vs. DPO
2. The Core Problem in Standard DPO
3. Proposed Solution: TUR-DPO Architecture
4. Mathematical Objective & Shaped Reward
5. Experimental Results
6. Conclusion & Future Work

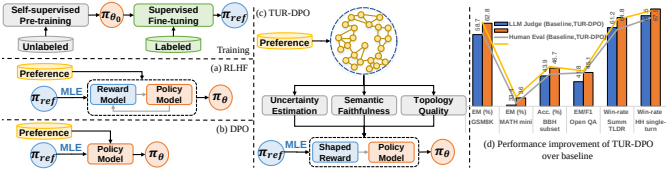
Background: RLHF vs. DPO

RLHF + PPO

- Highly effective for alignment.
- **Complex:** Requires online rollouts, a value head, reward shaping, and KL scheduling.
- Computationally expensive and memory-intensive.

Standard DPO

- Simpler and **RL-free**.
- Directly optimizes preferences without a separate reward model.
- **Limitation:** Treats preferences as *flat winner-vs-loser* signals.



(d) Performance improvement of TUR-DPO over baseline

The Core Problem

Key Gap in Standard DPO

Standard DPO cannot reward **how** an answer is derived or modulate learning when human preference labels are noisy or brittle.

Consequences:

- **Logical Leaps:** Models memorize correct answers without grasping the reasoning.
- **Hallucinations:** Lack of structural awareness leads to factually incorrect intermediate steps.
- **Overfitting to Noise:** No mechanism to scale down the loss when the human preference is highly subjective or incorrect.

Our Solution: TUR-DPO

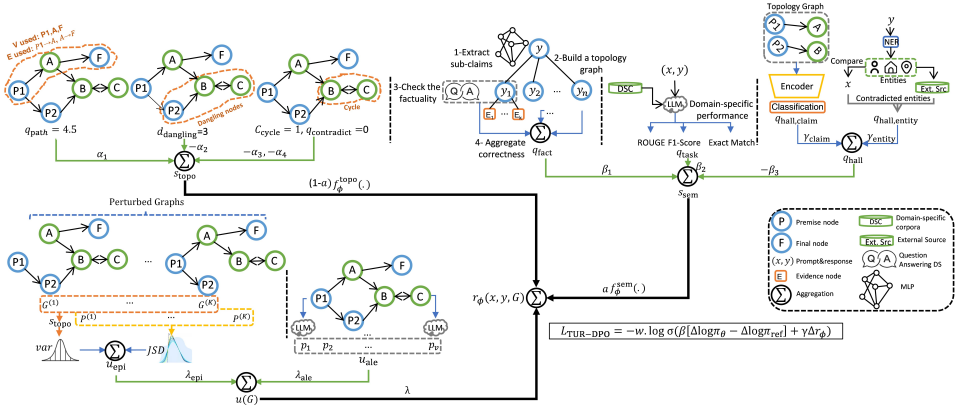
TUR-DPO augments DPO with structural verification and uncertainty awareness.

Three Core Pillars

1. **Lightweight reasoning topologies** (sub-claim directed graphs).
2. **Semantic & topology quality** scores.
3. **Calibrated uncertainty** weighting.

Why it matters: Forces the LM to learn *why* a response is better by verifying the intermediate reasoning steps for factual consistency. All within a simple **RL-free loop**.

Methodology: TUR-DPO Architecture



Methodology: Objective & Shaped Reward

Shaped Reward

Factorized over semantic, structural, and uncertainty signals:

$$r_\phi(x, y, G) = a\, f_\phi^{\mathsf{sem}}(s_{\mathsf{sem}}) + (1-a)\, f_\phi^{\mathsf{topo}}(s_{\mathsf{topo}}) - \lambda\, u(G)$$

TUR-DPO Loss

Uncertainty-weighted pairwise objective:

$$\mathcal{L}_{\text{TUR-DPO}} = -w \cdot \log \sigma\Big(\beta \left[\Delta \log \pi_\theta - \Delta \log \pi_{\mathsf{ref}}\right] + \gamma\, \Delta r_\phi\Big)$$

Mathematical Intuition & Robustness

Per-Pair Uncertainty Weight

$$w = \mathrm{clip}\left(\frac{\tau_w}{1+\bar{u}},\ w_{\min},\ 1\right), \qquad u(G) = \lambda_{\text{epi}}\, u_{\text{epi}} + \lambda_{\text{ale}}\, u_{\text{ale}}$$

Robustness Guarantee:

- w acts as an adaptive regularizer: shrinks when the topology is ambiguous to prevent overfitting to noisy preference pairs.
- $\gamma \Delta r_\phi$ dynamically expands the margin when reasoning is structurally sound and contracts it when support is weak.
- Safely handles out-of-distribution prompts and subjective tasks.

Experimental Results

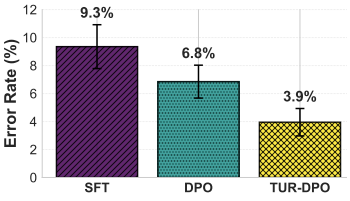
Human Evaluation (200 examples/task, double annotated)

Task	Metric	DPO	PPO	TUR-DPO
GSM8K	EM (%)	59.2	62.0	**63.1**
MATH	EM (%)	34.1	35.5	**36.4**
BBH	Acc (%)	44.5	46.0	**47.2**
QA	EM/F1	45.0	45.4	**45.7**
TLDR	Win (%)	60.5	63.7	**64.1**
HH	Win (%)	64.7	**67.9**	67.2

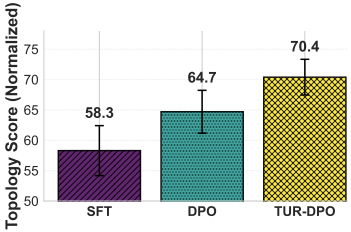
Key Takeaway

TUR-DPO matches or exceeds PPO on reasoning tasks **without online rollouts**. PPO retains a slight edge only on stylistic HH dialogue.

Conclusion & Future Work



Faithfulness (Lower is Better)



Coherence (Higher is Better)

Error Reduction (per 100 errors): Logical leaps: $31 \rightarrow 19$ Contradictions: $12 \rightarrow 7$ Hallucinations: $18 \rightarrow 13$

Future Work

- **Multi-modal Topologies:** Extending graphs to visual data.
- **Dynamic Scaling:** Learning adaptive λ parameters.

References

[1] Rafailov, R. et al. (2023).
Direct Preference Optimization: Your Language Model is Secretly a Reward Model.
Advances in Neural Information Processing Systems 36.

[2] Schulman, J. et al. (2017).
Proximal Policy Optimization Algorithms.
arXiv:1707.06347.

[3] Da, L. et al. (2025).
Understanding the Uncertainty of LLM Explanations: A Perspective Based on Reasoning Topology.
arXiv:2502.17026.

[4] Bai, Y. et al. (2022).
Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.
arXiv:2204.05862.

Thank You!

Questions & Discussion