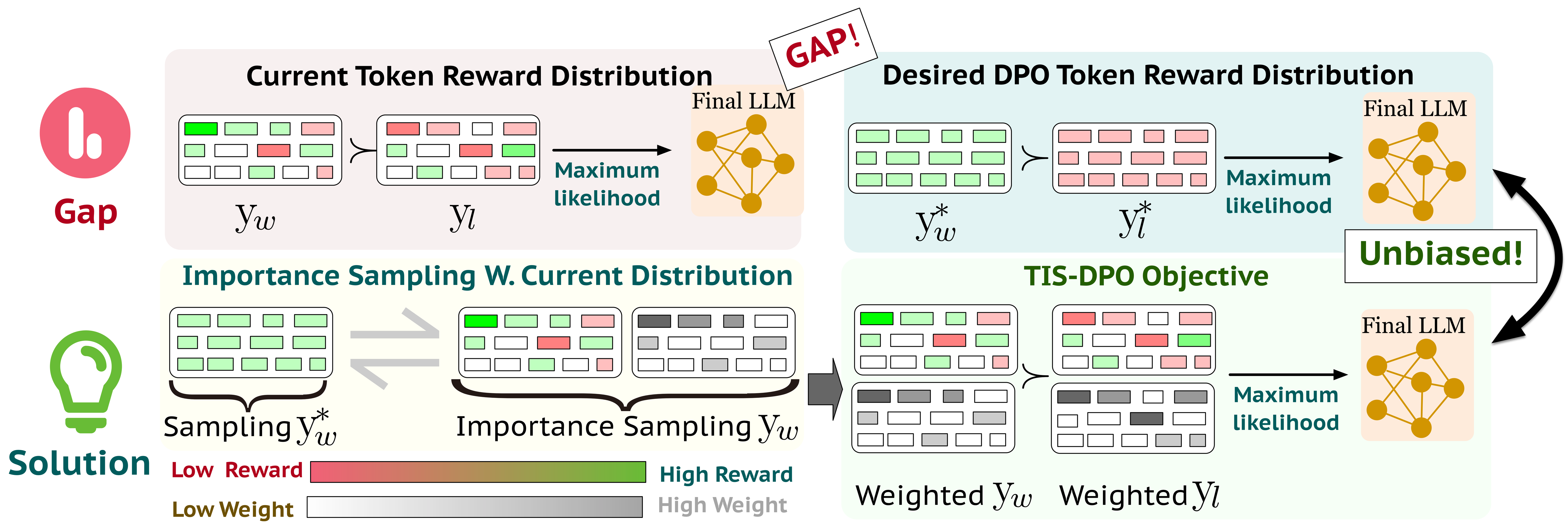
TIS-DPO: Token-level Importance Sampling for Direct Preference Optimization With Estimated Weights

Aiwei Liu[1], Haoping Bai[2], Zhiyun Lu[2], Yanchao Sun[2], Xiang Kong[2], Simon Wang[2], Jiulong Shan[2], Albin Madappally Jose[2], Xiaojiang Liu[2], Lijie Wen[1], Philip S. Yu[3], Meng Cao[2]

[1]Tsinghua University, [2]Apple, [3]University of Illinois at Chicago



Token Importance Estimation

Key Idea: Use contrastive LLMs to estimate token-level importance weights

Weight Estimation Formula:

$$w_t = k \cdot \exp(\mu \cdot \text{clamp}(\log \frac{\pi^+(y_t \mid x, y^{<t})}{\pi^-(y_t \mid x, y^{<t})}, L, U))$$

where π^+ favors high-reward tokens and π^- favors low-reward tokens

Three Methods for Contrastive LLM Construction:

- **TIS-DPO(P):** Prompt-based - Using contrastive prompts (e.g., "harmless" vs. "harmful")
- **TIS-DPO(S):** SFT-based - Fine-tuning separate models on winning vs. losing responses
- **TIS-DPO(D):** DPO-based - Training with DPO on normal and swapped preference pairs

Limitations of Direct Preference Optimization

DPO Objective: $\mathcal{L}_{\text{DPO}}$ (Maximizing likelihood ratio between winning and losing responses)

$$-\mathbb{E}_{(x,y_w,y_l)}\left[\log\sigma\left(\beta\sum_{i=1}^{n_w}\log\frac{\pi_\theta(y_w^i|x,y_w^{<i})}{\pi_{\text{ref}}(y_w^i|x,y_w^{<i})}-\beta\sum_{j=1}^{n_l}\log\frac{\pi_\theta(y_l^j|x,y_l^{<j})}{\pi_{\text{ref}}(y_l^j|x,y_l^{<j})}\right)\right]$$

Key Issues: Equal gradient for all tokens in winning/losing responses, regardless of their importance:

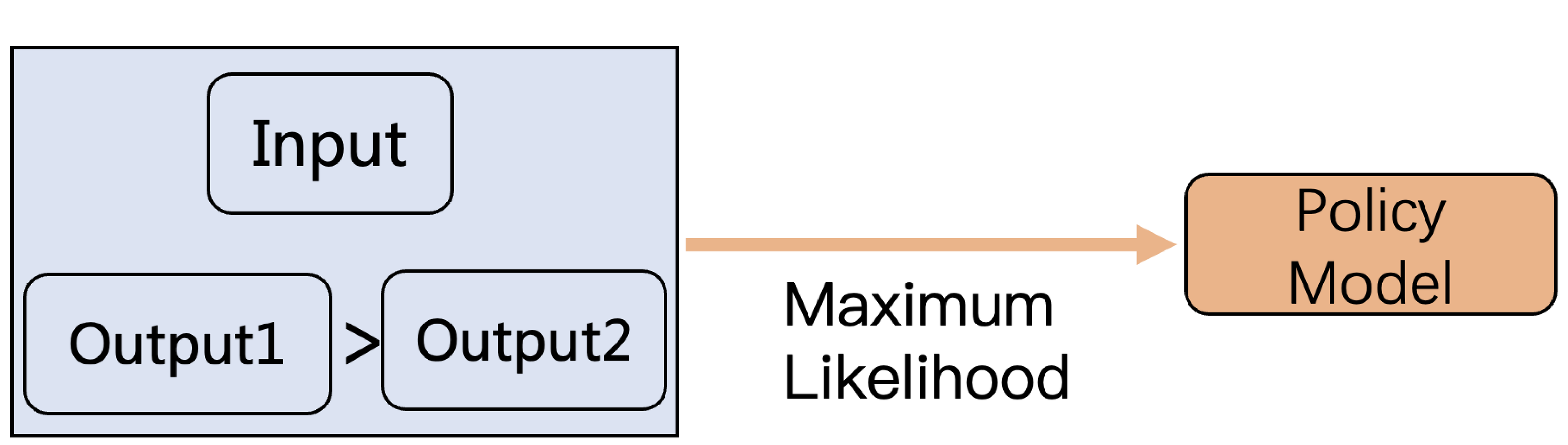
$$\frac{\partial\mathcal{L}_{\text{DPO}}}{\partial\log\pi_\theta(y_w^i)}=-\beta\cdot\sigma(-z)\quad\text{and}\quad\frac{\partial\mathcal{L}_{\text{DPO}}}{\partial\log\pi_\theta(y_l^j)}=\beta\cdot\sigma(-z)$$

where z is the log-ratio term, showing identical gradients for all tokens

Theoretical Bound on Optimization Stability: (Higher token reward variance leads to less stable optimization)

$$P(S_w\le S_l)\le\exp\left(-\frac{2n_wt^2}{(b_w-a_w)^2}\right)+\exp\left(-\frac{2n_lt^2}{(b_l-a_l)^2}\right)$$

where S_w, S_l: avg. rewards, $(b_w-a_w),(b_l-a_l)$: token reward ranges indicating variance



TIS-DPO Method

Our Solution: Assign importance weights to tokens based on their contribution to response quality

Step 1: Define optimal dataset $\mathcal{D}^*$ with consistent token rewards

$$\forall(x,y^{<t}),\quad\mathbb{E}_{y^t\sim\mathcal{D}^*(\cdot|x,y^{<t})}[r(y^t\mid x,y^{<t})]=R^*$$

Step 2: Derive token-level weights through importance sampling

$$D^*(x,y^{<t},y^t)=\frac{D(x,y^{<t},y^t)}{w(y^t\mid x,y^{<t})}$$

where $w(y^t\mid x,y^{<t})=k\cdot\exp(\mu r(y^t\mid x,y^{<t}))$ represents weights

Step 3: Reformulate Bradley-Terry model with token-level importance weights $P_{\text{BT}}(y_w\succ y_l\mid x)$

$$\sigma\left(\sum_{i=1}^{T_w}w_i^w\beta\log\frac{\pi_\theta(y_{w_i}\mid x,y_w^{<i})}{\pi_{\text{ref}}(y_{w_i}\mid x,y_w^{<i})}-\sum_{j=1}^{T_l}w_j^l\beta\log\frac{\pi_\theta(y_{l_j}\mid x,y_l^{<j})}{\pi_{\text{ref}}(y_{l_j}\mid x,y_l^{<j})}-\eta\right)$$

where η represents the difference in weighted KL divergence between winning and losing sequences

Definition of η:

$$\eta=\beta\left(\sum_{i=1}^{T_w}w_i^wD_{\text{KL}}(\pi_\theta\parallel\pi_{\text{ref}})_{x,y_w^{<i}}-\sum_{j=1}^{T_l}w_j^lD_{\text{KL}}(\pi_\theta\parallel\pi_{\text{ref}})_{x,y_l^{<j}}\right)$$

Final TIS-DPO Objective: $\mathcal{L}_{\text{TIS-DPO}}$

$$-\mathbb{E}_{(x,y_w,y_l)}\left[\log\sigma\left(\sum_{i=1}^{T_w}w_i^w\beta\log\frac{\pi_\theta(y_{w_i}|x,y_w^{<i})}{\pi_{\text{ref}}(y_{w_i}|x,y_w^{<i})}-\sum_{j=1}^{T_l}w_j^l\beta\log\frac{\pi_\theta(y_{l_j}|x,y_l^{<j})}{\pi_{\text{ref}}(y_{l_j}|x,y_l^{<j})}-\eta\right)\right]$$

TIS-DPO: Experiment Results

Datasets: PKU-SafeRLHF for harmlessness/helpfulness evaluation, with AdvBench, JailbreakBench, and Alpaca
Metrics: Llama-Guard safety detection, Beaver-Cost/Reward scoring, MT-bench, GPT-4 human preference
Hyperparameters: μ=±1, L=-0.5, U=1.5, k=1, β=0.1; trained for 1 epoch with RMSprop

Settings	*PKU-SafeRLHF*				
	Llama-Guard ↑	Harm. ↓	Help. ↑	MT ↑	Win ↑
LLaMA2-7B					
w. DPO	74.4%	5.6	7.9	4.1	-
w. PPO	78.7%	4.2	**8.1**	4.2	53.2%
w. IPO	74.8%	5.7	8.0	4.1	50.9%
w. TDPO	75.9%	4.6	8.0	4.1	52.4%
w. KTO	79.8%	4.1	8.0	4.0	58.3%
w. *TIS-DPO(P)*	75.9%	4.6	8.0	4.1	49.4%
w. *TIS-DPO(S)*	89.6%	3.2	7.8	**4.3**	66.7%
w. *TIS-DPO(D)*	**96.7%**	**0.1**	8.0	**4.3**	**79.3%**