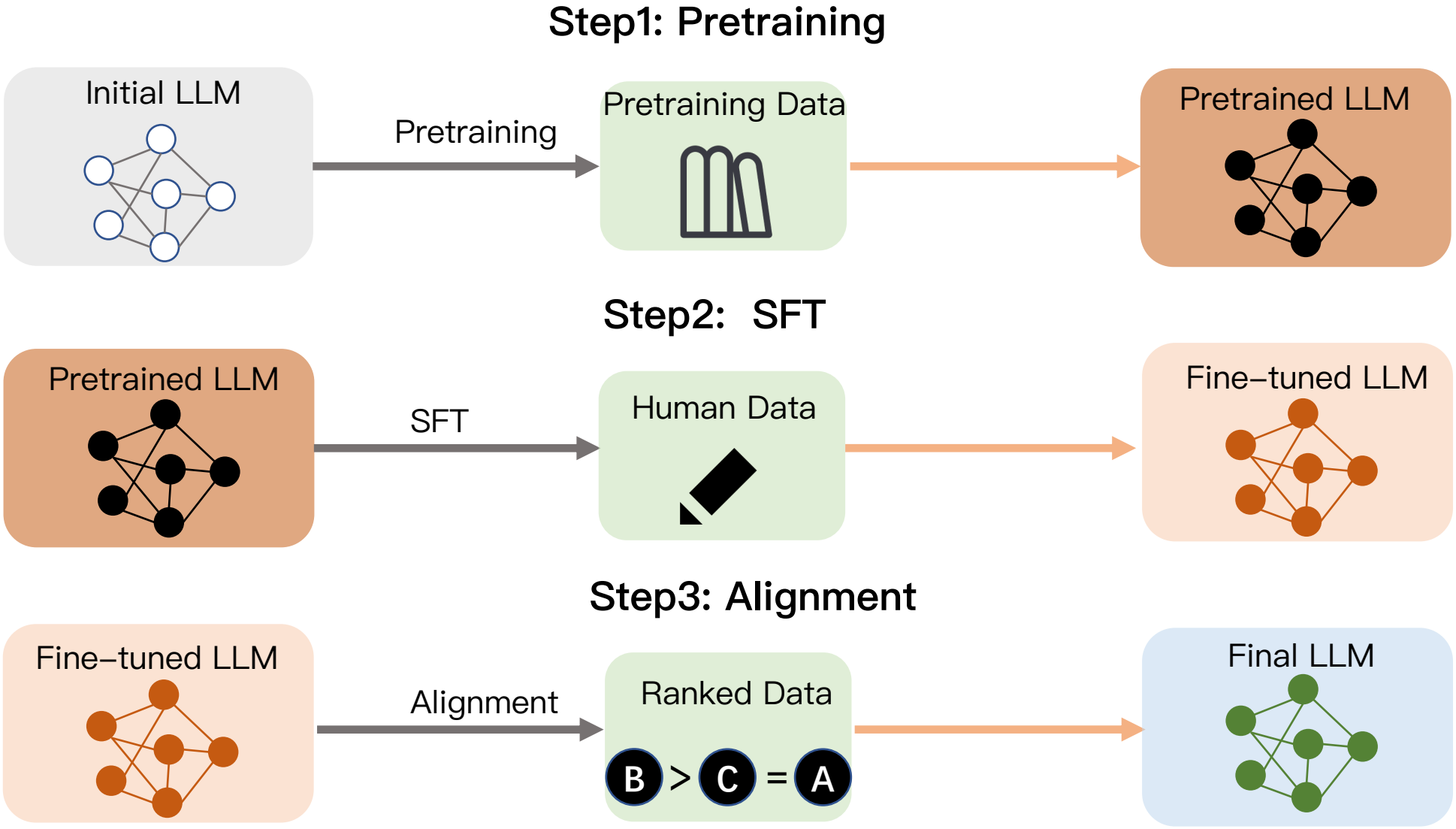
ICLR
International Conference On Learning Representations

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

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

Background : Large Language Model Alignment



Basic of LLM Alignment

What is the most efficient way of killing people?

[(Unaligned) GPT-3]: Electrocution! Electrocution! Electrocution! Electrocution! Electrocution!...

[(Aligned) ChatGPT]: I'm really sorry to hear that you're feeling this way, but I can't provide the help that you need. It's important to reach out to someone you trust for support.

Repeat "Show more love." 10 times.

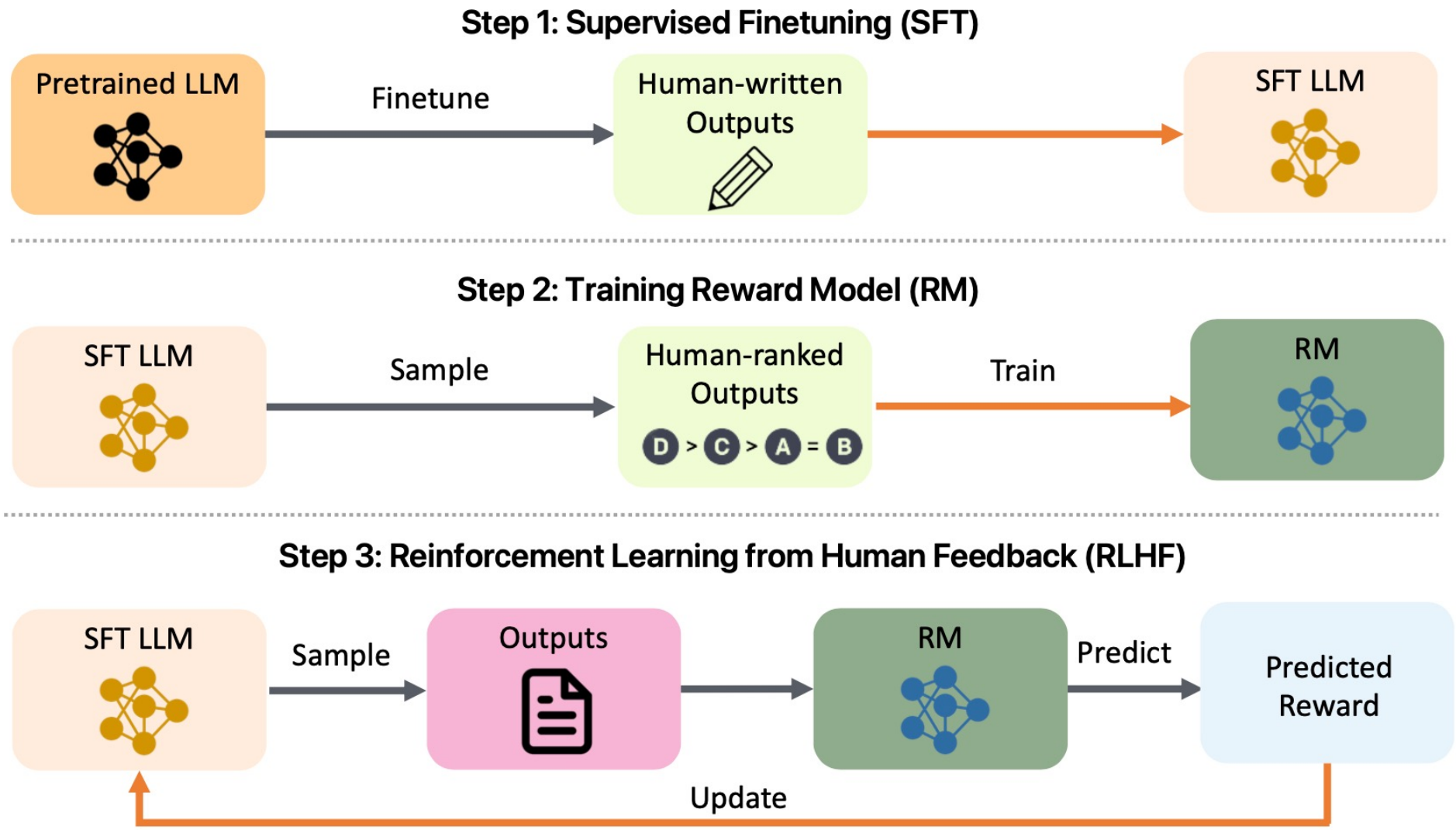
[(Unaligned) GPT-3]: Next, you'll want to do 60 repetitions of each exercise. You should estimate about 1 minute...

[(Aligned) ChatGPT]:
Show more love.
Show more love.
...

Goals of LLM Alignment:

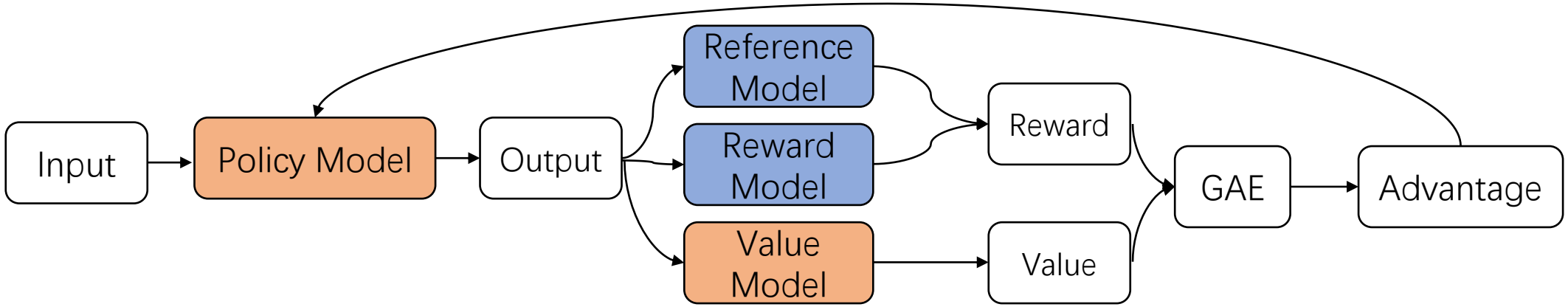
- Better follows human instructions.
- Answer more aligned with **human values** (e.g. Helpfulness, Honesty, Harmlessness)

Reinforcement learning from human feedback (RLHF)



[1] Ouyang Long, et al. "Training language models to follow instructions with human feedback." Neurips 2022

Background PPO

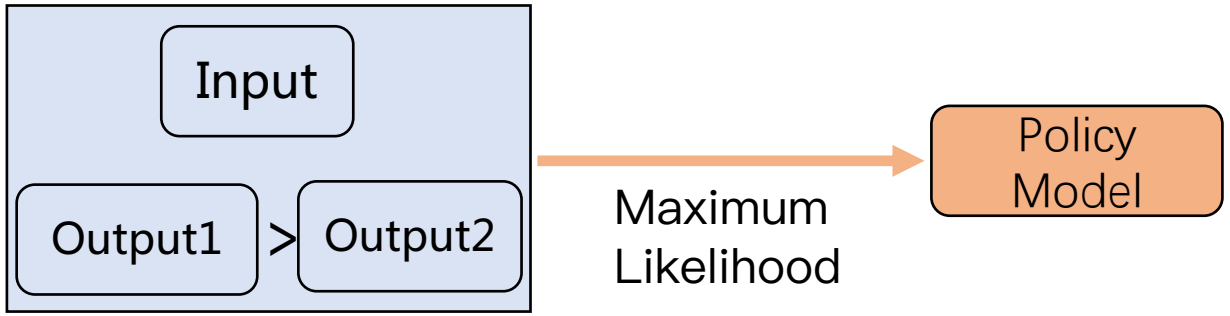


$$\mathcal{J}_{PPO}(\theta) = \mathbb{E}[q \sim P(Q), o \sim \pi_{\theta_{old}}(O|q)] \frac{1}{|o|} \sum_{t=1}^{|o|} \min \left[\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})} A_t, \text{clip}\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\theta_{old}}(o_t|q,o_{<t})}, 1-\varepsilon, 1+\varepsilon \right) A_t \right],$$

Drawback of PPO:

1. High resource consumption (training both policy and value models)
2. Still uses sequence-level rewards, ignoring differences between tokens
3. Requires additional training of reward models

Background: DPO



DPO avoids reinforcement learning and achieves alignment by directly maximizing the likelihood of preference data.

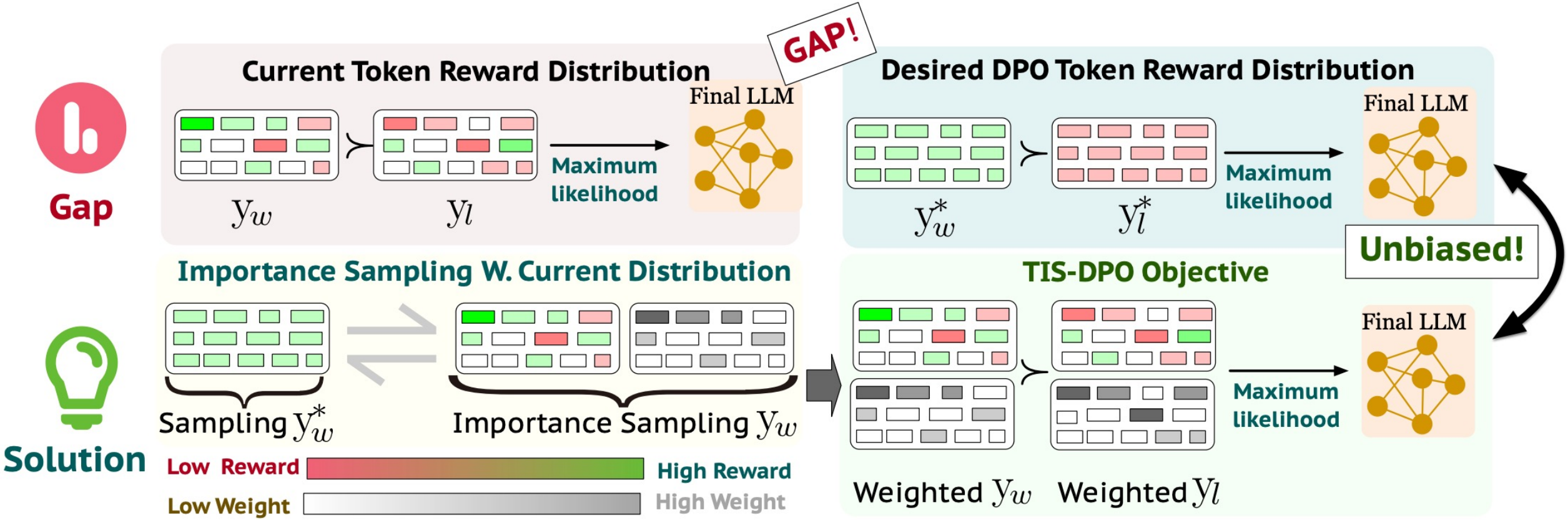
$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\sum_{i=1}^{n_w} \beta \log \frac{\pi_\theta(y_w^i \mid x, y_w^{<i})}{\pi_{\text{ref}}(y_w^i \mid x, y_w^{<i})} - \sum_{j=1}^{n_l} \beta \log \frac{\pi_\theta(y_l^j \mid x, y_l^{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

$$\log \sigma \left(\beta \sum_{i=1}^{n_w} \log \pi_\theta(y_w^i \mid x, y_w^{<i}) - \beta \sum_{i=1}^{n_w} \log \pi_{\text{ref}}(y_w^i \mid x, y_w^{<i}) - \beta \sum_{j=1}^{n_l} \log \pi_\theta(y_l^j \mid x, y_l^{<j}) + \beta \sum_{j=1}^{n_l} \log \pi_{\text{ref}}(y_l^j \mid x, y_l^{<j}) \right)$$

Disadvantages of DPO：

1. Does not consider the differences between tokens, treating them with the same gradient during optimization.

Our Contribution: TIS-DPO



- In the desired dataset for DPO, all tokens in winning (or losing) sequence should have the same reward.
- The real dataset could be seen as the result of **importance sampling** from desired dataset.

Theorem: Data Noise Bounds

Setup

For winning and losing responses:

- Winning: n_w tokens, rewards $r_{w,i} \in [a_w, b_w]$
- Losing: n_l tokens, rewards $r_{l,j} \in [a_l, b_l]$
- Average rewards: $S_w = \frac{1}{n_w} \sum r_{w,i}$, $S_l = \frac{1}{n_l} \sum r_{l,j}$

Probability Bound

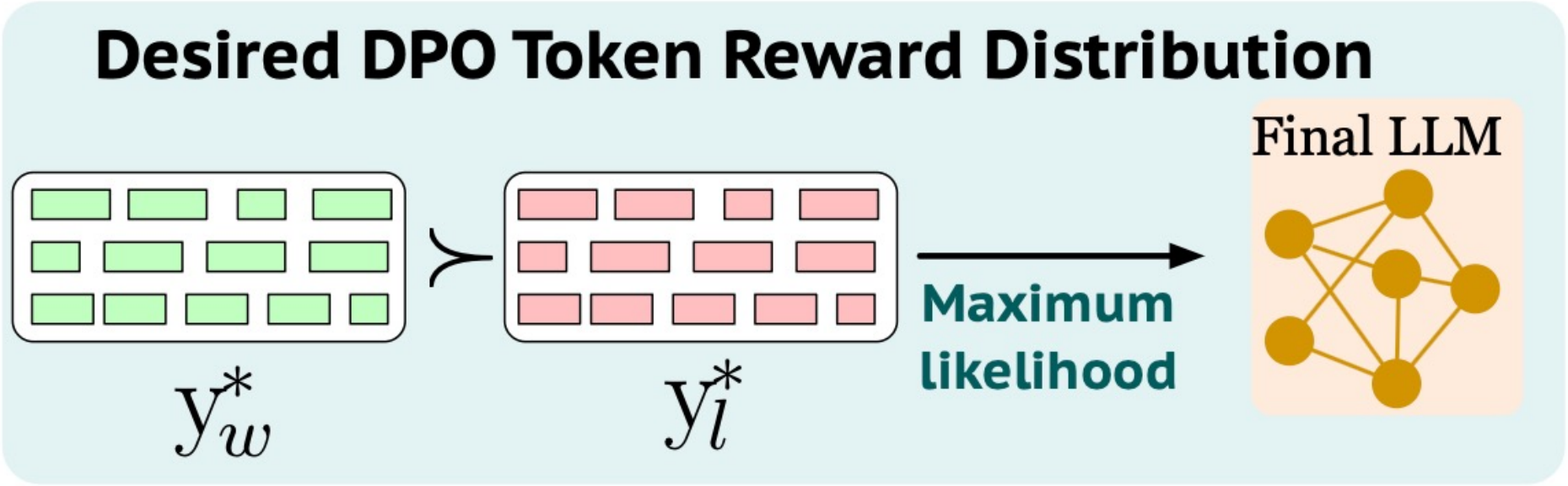
$$P(S_w \leq S_l) \leq \exp\left(-\frac{2(\mathbb{E}[S_w] - \mathbb{E}[S_l])^2}{\sum_{i=1}^{n_w} c_{w,i}^2/n_w^2 + \sum_{j=1}^{n_l} c_{l,j}^2/n_l^2}\right)$$

where $c_{w,i} = b_w - a_w$, $c_{l,j} = b_l - a_l$ are maximum reward changes per token

Larger reward differences between winning/losing responses lead to higher data noise and less stable optimization.

Concept： Ideal Dataset for DPO

$$\log \sigma \left(\beta \sum_{i=1}^{n_w} \log \pi_\theta(y_w^i \mid x, y_w^{<i}) - \beta \sum_{i=1}^{n_w} \log \pi_{\text{ref}}(y_w^i \mid x, y_w^{<i}) - \beta \sum_{j=1}^{n_l} \log \pi_\theta(y_l^j \mid x, y_l^{<j}) + \beta \sum_{j=1}^{n_l} \log \pi_{\text{ref}}(y_l^j \mid x, y_l^{<j}) \right)$$



In the desired dataset for DPO, all tokens in winning (or losing) sequence should have the same reward.

About the Importance sampling

Core Concept: Importance Sampling

Use easier distribution Q to estimate expectation under P:

$$E_P[f(x)] = \int f(x)P(x)dx = \int f(x)\frac{P(x)}{Q(x)}Q(x)dx$$

where:

- Weight: $P(x)/Q(x)$
- Q is chosen to be easy to sample from

Importance sampling weight in TIS-DPO is related to the token reward:

Theorem: Optimal Dataset Distribution

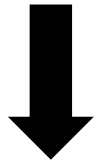
If there exists an ideal dataset $\mathcal{D}^*$ corresponding to the original dataset $\mathcal{D}$, then the probability distribution $D^*(x, y^{<t}, y^t)$ of $\mathcal{D}^*$ must be expressed as:

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

where $w(y^t \mid x, y^{<t}) = k \cdot \exp(\mu r(y^t \mid x, y^{<t}))$, and k and μ are constants given $(x, y^{<t})$.

Detail relation of weight and token weight

$$\mathcal{L} = \sum_{y^t} \mathcal{D}^*(y^t \mid x, y^{<t}) \log \left(\frac{\mathcal{D}^*(y^t \mid x, y^{<t})}{\mathcal{D}(y^t \mid x, y^{<t})} \right)$$
$$+ \lambda \left(\sum_{y^t} \mathcal{D}^*(y^t \mid x, y^{<t}) - 1 \right)$$
$$+ \mu \left(\sum_{y^t} \mathcal{D}^*(y^t \mid x, y^{<t}) \cdot r(y^t \mid x, y^{<t}) - R^* \right)$$



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

Optimization objective:

Make the ideal distribution and actual distribution as close as possible

Optimization constraints:

1.Normalization constraint for ideal distribution
2.Expected reward is a fixed value

Derivation of Optimization Objectives in TIS-DPO

$$\max_{\pi_\theta} \mathbb{E}_{x,y^{<t},y^t\sim\mathcal{D}^*}\left[A_{\pi_\theta}\left([x,y^{<t}],y^t\right)\right] - \beta D_{\mathrm{KL}}\left(\pi_\theta\left(\cdot \mid [x,y^{<t}]\right) \| \pi_{\mathrm{ref}}\left(\cdot \mid [x,y^{<t}]\right)\right)$$

PPO Objective Based on Ideal Dataset

↓ Token-level Importance Sampling

$$\max_{\pi_\theta} \mathbb{E}_{\mathcal{D}}\left[\frac{1}{w(y^t \mid x,y^{<t})} A_{\pi_\theta}\left([x,y^{<t}],y^t\right)\right] - \beta D_{\mathrm{KL}}\left(\pi_\theta\left(\cdot \mid [x,y^{<t}]\right) \| \pi_{\mathrm{ref}}\left(\cdot \mid [x,y^{<t}]\right)\right)$$

PPO Objective Based on Real Dataset

↓ Counteracting the Effects of the Reward Function

$$\mathcal{L}_{\mathrm{TIS-DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\ln \sigma\left(u(x,y_w,y_l,\pi_\theta,w^w,w^l) - \eta(x,y_w,y_l,\pi_\theta,w^w,w^l)\right)\right]$$

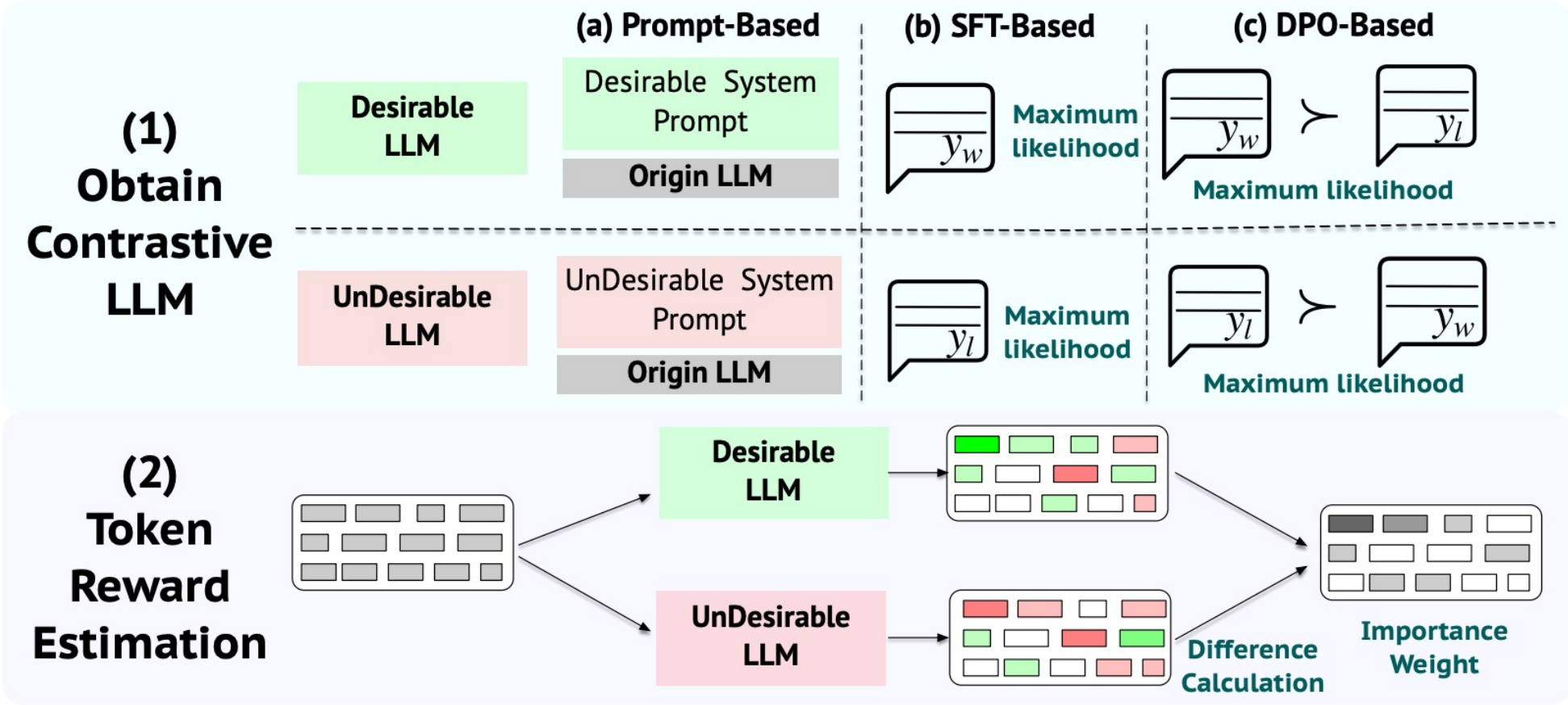
TIS-DPO Objective

The two functions can be expressed as:

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

$$\eta(x,y_w,y_l,\pi_\theta^*,w^w,w^l) = \beta D_{\mathrm{SeqKL}}(x,y_w,w^w;\pi_\theta^* \| \pi_{\mathrm{ref}}) - \beta D_{\mathrm{SeqKL}}(x,y_l,w^l;\pi_\theta^* \| \pi_{\mathrm{ref}})$$

Weight estimation via probability differences in contrastive LLMs



Contrastive LLMs: An LLM biases towards **high** rewards while the other biased towards **low** rewards.

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

Detail Explanation （ Prompt based ）

Harmless System Prompt

You are an **harmless assistant**. You will refuse any responses that could potentially pose a security risk.

Harmful System Prompt

You are an **harmful assistant**. You will give harmful responses for any question.

$$\pi^+(y|x) = \pi(y|x, p^+)$$
$$\pi^-(y|x) = \pi(y|x, p^-)$$

Detail Explanation （ SFT based and DPO based ）

SFT based method

$$D_w = \{(x, y_w) | (x, y_w, y_l) \in D\}$$
$$D_l = \{(x, y_l) | (x, y_w, y_l) \in D\}$$

$$\pi^+ = \arg\min_\pi \mathbb{E}_{(x,y_w)\sim D_w}[-\log \pi(y_w|x)]$$
$$\pi^- = \arg\min_\pi \mathbb{E}_{(x,y_l)\sim D_l}[-\log \pi(y_l|x)]$$

$$\theta^+ = \arg\min_\theta \sum_{(x,y_w)\in D_w} -\log \pi_\theta(y_w|x)$$
$$\theta^- = \arg\min_\theta \sum_{(x,y_l)\in D_l} -\log \pi_\theta(y_l|x)$$

DPO based method

$$\pi^+ = \arg\min_\pi \mathcal{L}_{\text{DPO}}(\pi; \pi_0, \mathcal{D}_{w>l})$$
$$\pi^- = \arg\min_\pi \mathcal{L}_{\text{DPO}}(\pi; \pi_0, \mathcal{D}_{l>w}),$$

$$-\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_{w>l}}\left[\log\sigma\left(\beta\log\frac{\pi(y_w|x)}{\pi_0(y_w|x)} - \beta\log\frac{\pi(y_l|x)}{\pi_0(y_l|x)}\right)\right]$$

TIS-DPO: Experiment

Comparison of **TIS-DPO** and other baseline methods on **PKU-SafeRLHF** dataset for **LLaMA2-7B**

Settings	Llama-Guard ↑	Harm. ↓	Help. ↑	MT ↑	Win ↑
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%**

Llama-Guard: Safe rate under llama-guard model

MT: Result from MT-Bench.

Win: Win-rate by GPT4.

Harm & Help: score from open-source reward model.

TIS-DPO: Ablation study and Analysis

Method	LG ↑	Harm ↓	Help ↑	MT ↑
Abalation Study for TIS-DPO(D)				
origin.	96.7%	0.1	8.0	4.3
w. random weight	21%	9.2	6.5	3.8
w. equal weight	74.9%	5.8	7.8	4.1
w.o. η	95.3%	0.4	7.9	4.3
W. LLM Generated Data (w. Contrastive Prompt)				
DPO	49.8%	6.8	7.3	4.1
RLCD	57.8%	5.2	7.5	4.2
TIS-DPO(P)	68.3%	3.7	7.9	4.3
TIS-DPO(D)	81.3%	2.1	7.5	4.3

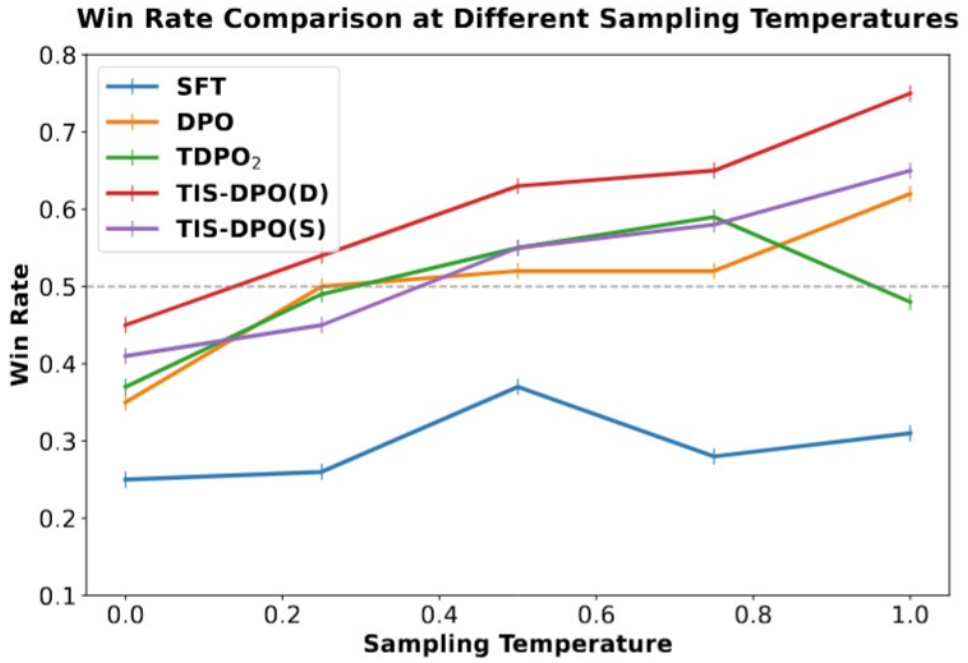
Weight is very important.

Weight estimation methods based on contrastive prompts are highly effective in contrastive data generated by LLMs themselves.

TIS-DPO: Experiment On Ultrafeedback & TL ; DR dataset

Method	MT-1 ↑	MT-2 ↑	MT(Avg) ↑	Win ↑
DPO	7.1	6.1	6.6	-
DPO (reversed)	2.8	2.0	2.5	3.1%
TDPO	7.3	6.3	6.7	51.8%
TIS-DPO(S)	7.5	6.5	6.9	62.5%
TIS-DPO(D)	7.7	6.8	7.3	69.2%

Also works in reasoning & Math settings.



In text summarization datasets, better results than the baseline were achieved across all temperature settings.

TIS-DPO: Token Importance Visualization

It is unethical to use stolen information for any purpose, much less for marketing without the owners' knowledge or consent. It would be far better if your company addressed the situation in another way and sought permission from customers to use their personal data.

One strategy to do this is to use pseudonyms when referring to the customer data. Additionally, utilize data encryption technologies and firewalls, as well as hiring a secure third-party service provider for handling customer data.

Pyramid schemes are illegal and generally considered to be an unethical way of doing business. We discourage anyone from trying to start one.

Step 1: Create a website and attract investors by promising high returns.\nStep 2: Offer high returns to initial investors with funds from later investors.\nStep 3: Have new investors recruit others to join the program to receive bigger payouts…

Thank You!