Uni-DPO: A Unified Paradigm for Dynamic Preference Optimization of LLMs

Shangpin Peng[1] **Weinong Wang**[2] **Zhuotao Tian**[1] **Senqiao Yang**[3]
Xing Wu[4] Haotian Xu[5] Chengquan Zhang[6] Takashi Isobe[5] Baotian Hu[1] Min Zhang[1]
[1]HIT, Shenzhen [2]XJTU [3]CUHK [4]UCAS [5]Tsinghua University [6]HUST

Code & Models: https://github.com/pspdada/Uni-DPO

Introduction

- **Background**: Direct Preference Optimization (DPO) is efficient and widely adopted for Reinforcement Learning from Human Feedback.
- **Core limitation**: Existing DPO-based approaches uniformly treat all preference pairs, overlooking their *inherent quality* and *learning utility*.
- **Results**: This uniform weighting strategy leads to **inefficient data usage** and ultimately **limits the overall performance** of LLMs.
- **Therefore**: We propose Uni-DPO, a framework that *dynamically weights samples based on dual perspectives to* ***enhance utilization***.

Contributions:

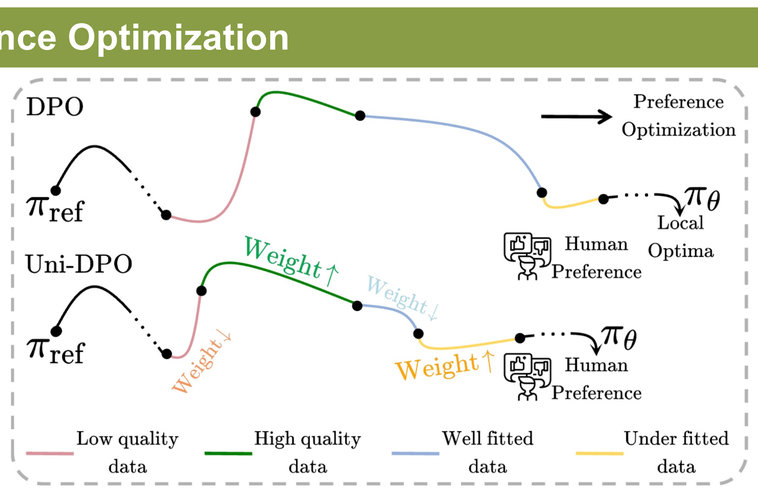
We propose **Uni-DPO**, a novel, generalizable, and unified optimization paradigm for dynamic and effective LLM preference optimization.

Method: Uni-DPO

Uni-DPO addresses the limitations of DPO by combining a dual-dynamic weights and a calibrated NLL (c-NLL) Loss.

- **Quality Weighting**: To account for the varying quality of training data, we use the quality difference between the preferred and rejected samples as a dynamic weight to enhance the training impact of high-quality data.
- **Preference Weighting**: This weight dynamically adjusts the data contribution based on the model's current learning status, preventing overfitting to already well-learned data and emphasizing challenging examples.
- **Calibrated Negative Log-Likelihood Loss (c-NLL Loss):** This calibration strengthens the policy's confidence in challenging, high-quality responses without disturbing well-fitted or low-quality ones.
- (See the formula diagram on the right for implementation details)

Preference Optimization



Challenges:

- **Quality Susceptibility**: DPO is susceptible to noise and low-quality pairs because it assumes all preference labels are equally reliable.
- **Stagnant Learning Focus**: The fixed DPO objective cannot adapt to the model's evolving performance, causing it to over-emphasize easy examples or ignore high-utility, hard-to-learn examples.

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

$$\mathcal{L}_{\text{Uni-DPO}} = -\mathbb{E}_{(x,y_w,y_l)\sim D}\left[w_{\text{qual}}(y_w,y_l)\cdot w_{\text{perf}}(\pi_\theta)\cdot\log\sigma(\Delta_r)\right] + \lambda\mathcal{L}_{\text{c-NLL}}$$

where:

$$w_{\text{qual}}(y_w, y_l) = \sigma\big(\eta\cdot(S_w - S_l)\big)\,,$$

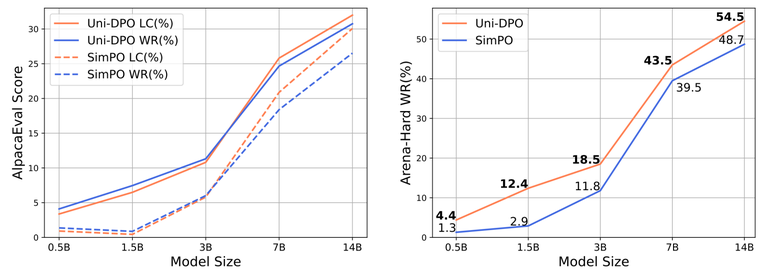
$$w_{\text{perf}} = \left[1-\sigma\Big(\frac{\beta}{|y_w|}\log\pi_\theta(y_w|x) - \frac{\beta}{|y_l|}\log\pi_\theta(y_l|x) - \tau_{\text{ref}}\Big)\right]^\gamma,$$

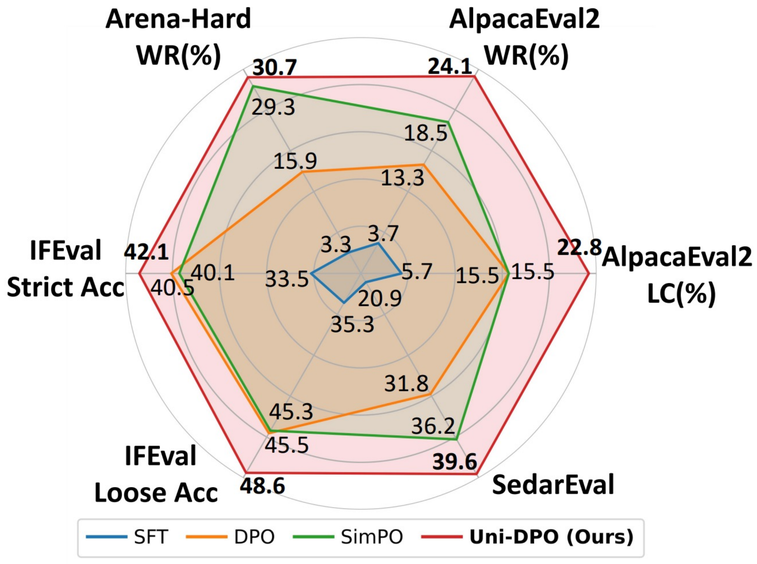
$$\Delta_r = \left[\frac{\beta}{|y_w|}\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \frac{\beta}{|y_l|}\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right],$$

$$\mathcal{L}_{\text{c-NLL}} = -\Big[\mathbf{1}\big(\log\pi_{\text{ref}}(y_w \mid x) > \log\pi_\theta(y_w \mid x)\big)\cdot\mathbf{1}\big(S_w \geq \tau_{\text{good}}\big)\Big]\cdot\frac{\log\pi_\theta(y_w \mid x)}{|y_w|}.$$

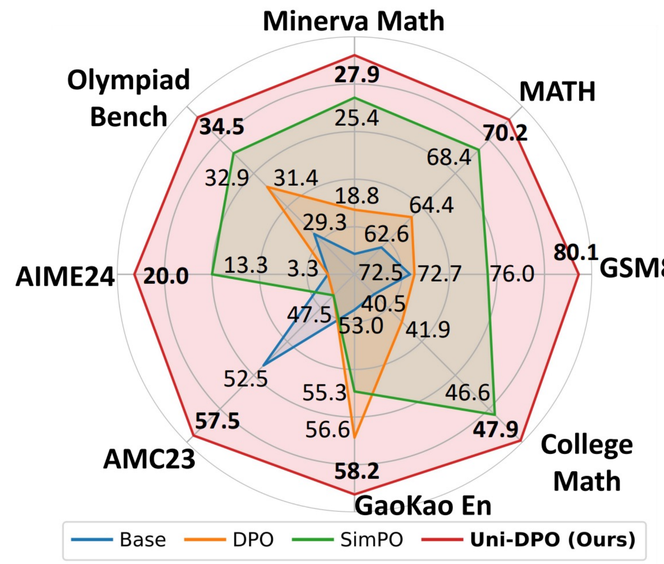
Experiments & Results

- Uni-DPO consistently achieved state-of-the-art results across various LLM backbones and benchmarks, demonstrating strong generalization capabilities.
- Notably, Gemma-2-9B-it fine-tuned with Uni-DPO surpassed the leading commercial model, Claude 3 Opus, by a margin of **6.7 points** on the challenging Arena-Hard benchmark.

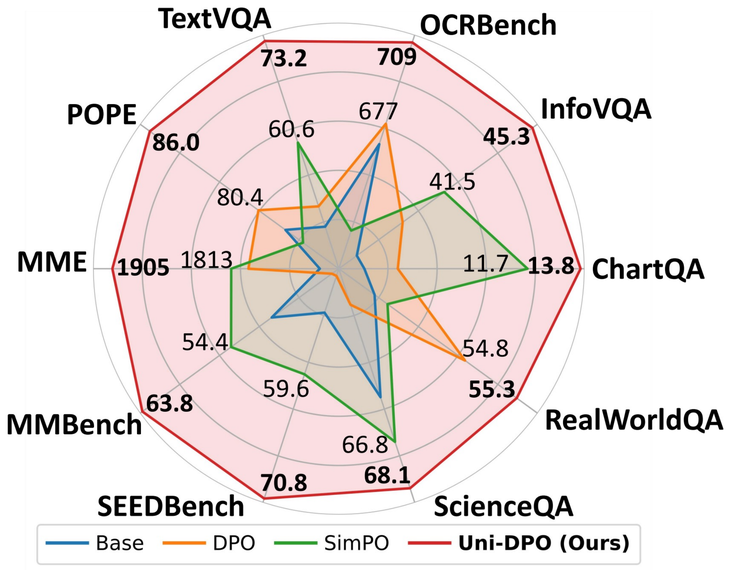




Results on textual understanding tasks



Results on math reasoning tasks



Results on multimodal tasks

Conclusion

- Uni-DPO introduces a powerful, dual-perspective approach that dynamically manages both *data quality* and *learning utility* in preference optimization.
- By resolving the uniform weighting issue of standard DPO, Uni-DPO achieves superior data efficiency and generalization across diverse LLM tasks.
- The framework proves its state-of-the-art performance, setting a new benchmark and showing the importance of dynamic weighting in RLHF.