SimPER: A Minimalist Approach to Preference Alignment without Hyperparameters

Teng Xiao* **Yige Yuan*** Zhengyu Chen Mingxiao Li Shangsong Liang Zhaochun Ren Vasant G Honavar

PennState

Prior Alignment Approaches

Using reinforcement learning fine-tunes the policy by optimizing the reward model based human preferences.

(Ouyang et al., 2022; Christiano et al., 2017; Schulman et al., 2017)

$$\max_{\pi_\theta} \mathbb{E}_{\mathbf{y}\sim\pi_\theta(\mathbf{y}|\mathbf{x})}\left[r(\mathbf{x},\mathbf{y})\right] - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi_\theta(\mathbf{y}\mid\mathbf{x})\|\pi_{\mathrm{ref}}(\mathbf{y}\mid\mathbf{x})\right]$$

... but training are **expensive**

Other approaches don't need training reward models e.g., by directly optimizing policy (DPO & SimPO).

(Rafailov et al., 2024; Azar et al., 2024; Meng et al., 2024; Ethayarajh., 2024)

$$\mathcal{L}_{\mathrm{DPO}}(\theta;\mathcal{D}) = \mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}}\left[-\log\sigma\left(\beta\log\frac{\pi_\theta(\mathbf{y}_w\mid\mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}_w\mid\mathbf{x})} - \beta\log\frac{\pi_\theta(\mathbf{y}_l\mid\mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}_l\mid\mathbf{x})}\right)\right]$$

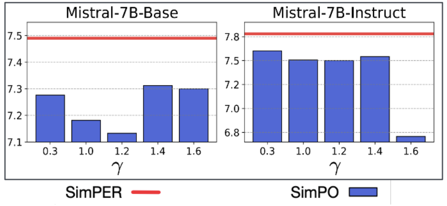
$$\mathcal{L}_{\mathrm{SimPO}}(\theta;\mathcal{D}) = \mathbb{E}_{(\mathbf{x},\mathbf{y}_w,\mathbf{y}_l)\sim\mathcal{D}}\left[-\log\sigma\left(\frac{\beta}{|\mathbf{y}_w|}\log\pi_\theta(\mathbf{y}_w\mid\mathbf{x}) - \frac{\beta}{|\mathbf{y}_l|}\log\pi_\theta(\mathbf{y}_l\mid\mathbf{x}) - \gamma\right)\right]$$

... but require expensive **hyperparameter** tuning

Problem: Hyperparameters

Current alignment methods are **highly sensitive** to **hyper-parameters**, which must be carefully tuned.

Method	Hyperparameters	#Hyperparameters	w/o Reference Model
DPO	β	1	✗
IPO	β	1	✗
KTO	$\lambda_l, \lambda_w, \beta$	3	✗
CPO	λ, β	2	✓
SLiC	δ, λ	2	✓
SimPO	γ, β	2	✓
SimPER	-	0	✓



SimPER:

Simple alignment with **Per**plexity optimization

tl;dr: a **simple** algorithm (**SimPER**) for preference alignment on LLMs **without hyperparameters**

code: https://github.com/tengxiao1/SimPER

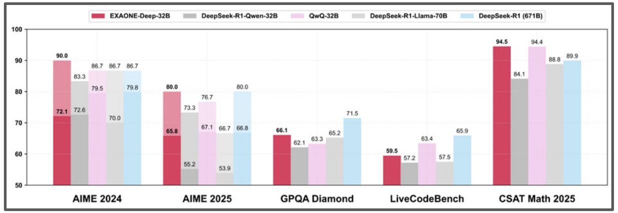
$$\mathcal{L}_{\mathrm{SimPER}}(\theta;\mathcal{D}) = -\mathrm{Perplexity}^{-1}(\mathbf{y}_w\mid\mathbf{x}) + \mathrm{Perplexity}^{-1}(\mathbf{y}_l\mid\mathbf{x})$$
$$= -\exp\left(\frac{1}{|\mathbf{y}_w|}\log\pi_\theta(\mathbf{y}_w\mid\mathbf{x})\right) + \exp\left(\frac{1}{|\mathbf{y}_l|}\log\pi_\theta(\mathbf{y}_l\mid\mathbf{x})\right)$$

How does SimPER perform?

SimPER achieves the **best ranking** across different models over 10 benchmarks, **without any hyperparameters.**

	Method	MMLU-PRO	IFEval	BBH	GPQA	MUSR	MATH	GSM8K	ARC	TruthfulQA	Winograd	Avg. Rank
Mistral-7B Base	DPO	26.73	10.49	43.27	28.44	43.65	1.36	21.76	61.26	53.06	76.80	4.7
	SLiC	26.52	12.45	42.33	27.93	33.74	1.38	33.74	55.38	48.36	77.35	5.0
	IPO	25.87	11.52	40.59	28.15	42.15	1.25	27.14	60.84	45.44	**77.58**	5.4
	KTO	27.51	12.03	43.66	29.45	43.17	2.34	**38.51**	62.37	**56.60**	77.27	2.5
	CPO	27.04	13.32	42.05	28.45	42.15	2.15	33.06	57.00	47.07	76.48	4.5
	SimPO	27.13	10.63	42.94	29.03	39.68	2.49	22.21	62.63	50.68	77.54	3.8
	SimPER	**27.84**	**15.83**	**43.99**	**30.12**	**43.95**	**2.57**	33.02	**63.50**	53.64	76.25	**2.0**
LLama3-8B Base	DPO	31.58	33.61	47.80	32.23	40.48	4.53	38.67	64.42	53.48	76.80	4.2
	SLiC	31.11	32.37	46.53	33.29	40.55	3.92	48.82	61.43	54.95	**77.27**	4.5
	IPO	30.18	31.52	46.78	32.61	39.58	4.02	22.67	62.88	54.20	72.22	6.4
	KTO	31.16	37.10	47.98	33.72	40.21	4.14	38.97	63.14	55.76	76.09	4.0
	CPO	30.95	38.57	47.17	33.15	41.59	4.25	46.93	61.69	54.29	76.16	4.2
	SimPO	31.61	37.55	48.38	33.22	40.08	4.23	31.54	65.19	59.46	76.32	3.4
	SimPER	**31.99**	**41.78**	**48.62**	**33.80**	**46.03**	**4.61**	**51.02**	**67.06**	**62.59**	76.24	**1.3**

SimPER is used for **EXAONE Deep 32B** at LG AI Research, resulting in **exciting reasoning performance**.



Why does SimPER work?

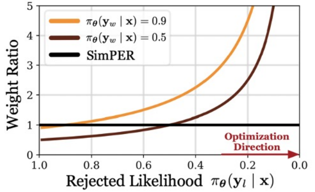
SimPER **balances gradients** by removing the log term, mitigating gradient dominance issue of negative samples.

$$\boxed{\frac{\pi_\theta(\mathbf{y}_w\mid\mathbf{x})}{\pi_\theta(\mathbf{y}_l\mid\mathbf{x})}}\cdot\frac{\nabla_\theta\pi_\theta(\mathbf{y}_l\mid\mathbf{x})}{\nabla_\theta\pi_\theta(\mathbf{y}_w\mid\mathbf{x})}$$

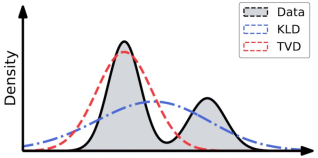
SimPO

$$\frac{\nabla_\theta\pi_\theta(\mathbf{y}_l\mid\mathbf{x})}{\nabla_\theta\pi_\theta(\mathbf{y}_w\mid\mathbf{x})}$$

SimPER



SimPER optimizes the **Total Variation Distance (TVD)**, offering a **mode-seeking** advantage over SFT.



$$\min_\theta \mathcal{L}_{\mathrm{SFT}} \Rightarrow \min_\theta \mathrm{KL}(\pi_{\mathrm{chosen}}(\mathbf{y}\mid\mathbf{x})\|\pi_\theta(\mathbf{y}\mid\mathbf{x})) = \sum_{\mathbf{y}\in\mathcal{Y}}\pi_{\mathrm{chosen}}(\mathbf{y}\mid\mathbf{x})\log\frac{\pi_{\mathrm{chosen}}(\mathbf{y}\mid\mathbf{x})}{\pi_\theta(\mathbf{y}\mid\mathbf{x})} \quad (12)$$

$$\min_\theta \mathcal{L}_{\mathrm{SimPER}} \Rightarrow \min_\theta \mathrm{TV}(\pi_{\mathrm{chosen}}(\mathbf{y}\mid\mathbf{x})\|\pi_\theta(\mathbf{y}\mid\mathbf{x})) = \frac{1}{2}\sum_{\mathbf{y}\in\mathcal{Y}}\left|\pi_{\mathrm{chosen}}(\mathbf{y}\mid\mathbf{x}) - \pi_\theta(\mathbf{y}\mid\mathbf{x})\right| \quad (13)$$

SimPER exhibits the **least decline** in chosen likelihoods while maintaining the **largest margin** between chosen