Noise Contrastive Alignment of Language Models with *Explicit* Rewards

Huayu Chen, Guande He, Lifan Yuan, Ganqu Cui, Hang Su, Jun Zhu

Tsinghua University

We propose a general LLM alignment framework that can:

1) Address the **chosen likelihood decrease** problem of DPO.

2) Handle alignment dataset labeled by **scalar rewards**.

3) Unifies contrastive learning (NCE) and LLM alignment theories.

4) Subsumes **DPO as a special case** of InfoNCE-based methods.

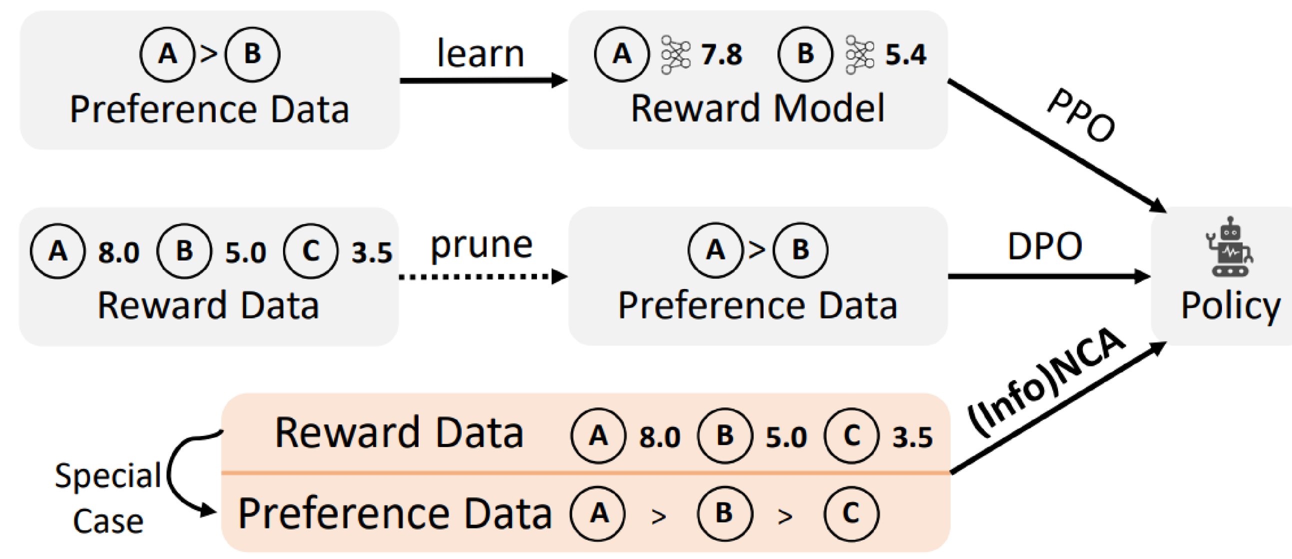


Figure 1: InfoNCA/NCA allows direct LM optimization for both reward and preference data.

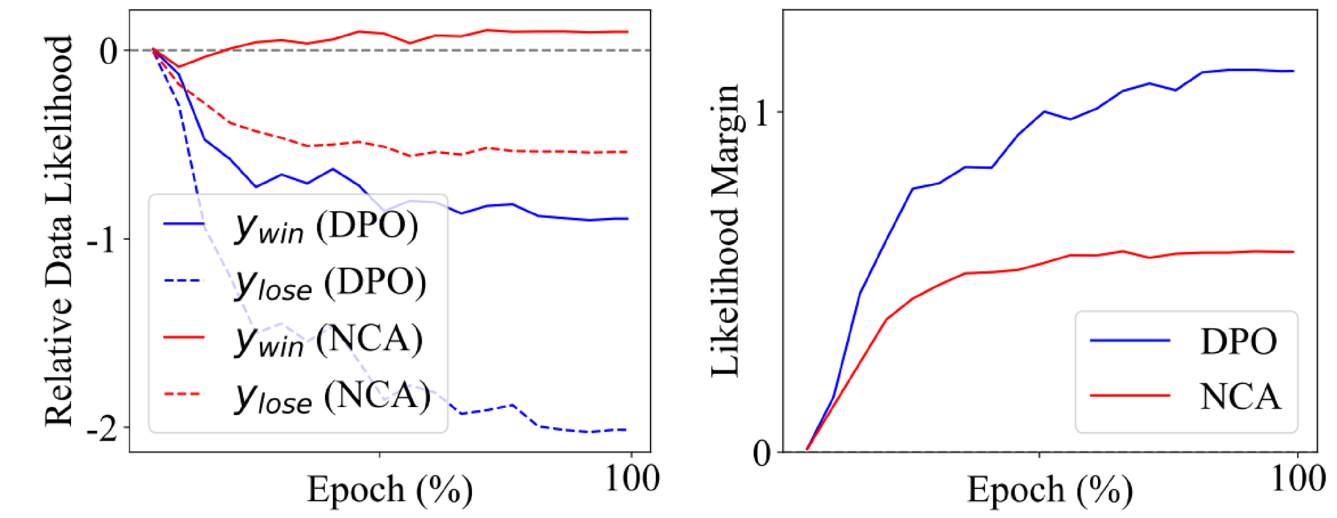
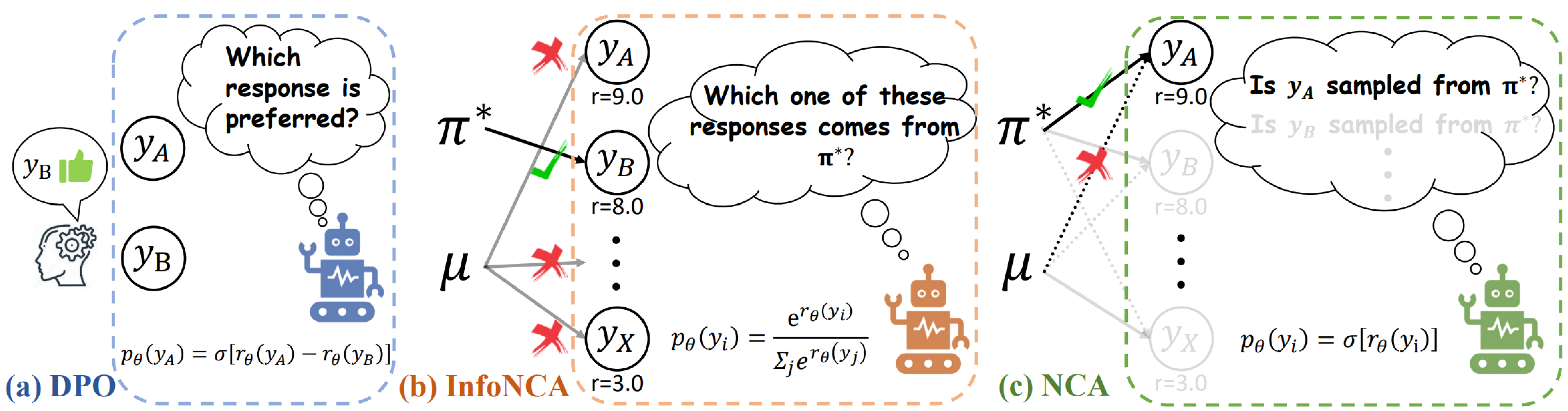


Figure 2: Pairwise NCA prevents chosen likelihood from decreasing while DPO cannot.

Method: InfoNCA and NCA methods for both reward&preference alignment.

Alignment Method	**InfoNCA** (Sec. 3)	**NCA** (Sec. 4)
Modeling Target	$\pi^*(y\|x) \propto \mu(y\|x)e^{r(x,y)/\alpha}$	
Model Definition	$\pi_\theta(y\|x) \propto \mu(y\|x)e^{r_\theta(x,y)}$	$\pi_\theta(y\|x) = \mu(y\|x)e^{r_\theta(x,y)}$
Reward Dataset	$x \rightarrow \{y_i, r_i\}_{1:K}$	
Loss (K>1,α>0)	$-\sum_{i=1}^{K}\left[\frac{e^{r_i/\alpha}}{\sum_j e^{r_j/\alpha}} \log \frac{e^{r_\theta(x,y_i)}}{\sum_j e^{r_\theta(x,y_j)}}\right]$	$-\sum_{i=1}^{K}\left[\frac{e^{r_i/\alpha}}{\sum_j e^{r_j/\alpha}} \log\sigma(r_\theta(x,y_i)) + \frac{1}{K}\log\sigma(-r_\theta(x,y_i))\right]$
Preference Dataset	$x \rightarrow \{y_w > y_l\}$	
Loss (K=2, $\alpha\rightarrow$0)	$-\log\sigma(r_\theta(x,y_w) - r_\theta(x,y_l))$(DPO)	$-\log\sigma(r_\theta(x,y_w)) - \frac{1}{2}\sum_{y\in\{y_w,y_l\}}\log\sigma(-r_\theta(x,y))$
Loss Type	InfoNCE loss [24]	NCE loss [14]
Optimizing Target	*relative* value of log likelihood ratio	*absolute* value of log likelihood ratio
Optimal $r_{\theta^*}(x,y)$	$r(x,y)/\alpha + C(x)$	$r(x,y)/\alpha - \log \mathbb{E}_{\mu(y\|x)} e^{r(x,y)/\alpha}$
$r_{\theta^*}(x, y_{\text{best}}) \geq 0$?	not guaranteed	✔



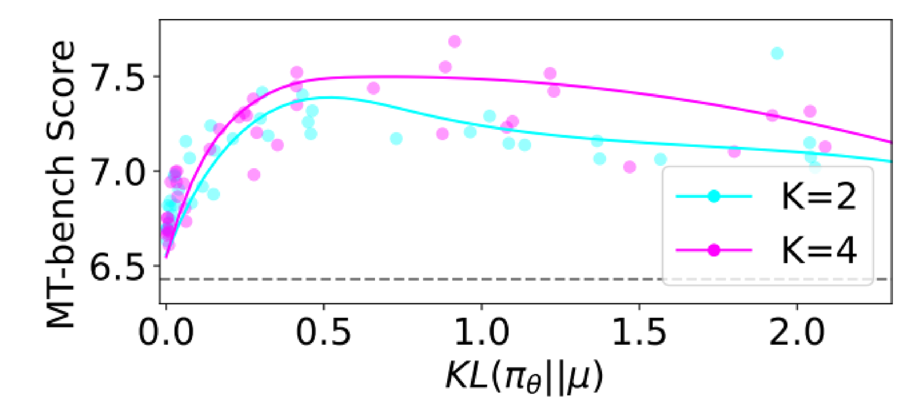
Experimental Findings:

1) Reward information is helpful and useful. Do not throw them away!

	Name	Annotation Type	MT-bench	AlpacaEval	Win vs. DPO
Baseline	Mixtral-7B-sft	SFT Data	6.45	85.20	-
	+KTO [11]	Preference	7.12	<u>91.93</u>	-
	+IPO [1]	Preference	7.45	90.62	-
	+DPO (Zephyr-β)	Preference	7.34	90.60	50.0
	+DPO$\times$3	Preference	7.22	91.60	<u>58.1</u>
	+DPO$\times C_4^2$	Preference	7.38	90.29	48.1
Ours	**+InfoNCA**	Reward	**7.63**	**92.35**	56.9
	+NCA	Reward	<u>7.52</u>	90.31	**59.4**

2) Suboptimal responses are also important for LLM alignment.

Method	K=2	K=3	K=4
InfoNCA (MT-bench)	73.8	75.9	76.3
InfoNCA (Alpaca)	90.7	90.2	92.4
NCA (MT-bench)	73.2	73.3	75.2
NCA (Alpaca)	89.9	90.3	90.3
Average	**81.9**	**82.4**	**83.5**



3) NCA is extremely helpful in reasoning tasks like math and coding.

Model	Reasoning BBH (CoT)	Coding LeetCode	Coding HumanEval	Math GSMPLUS	Math MATH	Math TheoremQA	Math SVAMP	Math ASDiv	Avg.
Mixtral-7B-SFT	60.9	3.3	28.1	28.5	5.8	7.0	26.9	35.8	24.5
+ DPO	**61.7**	2.2 ↓	**31.7**	12.1 ↓	6.4	9.8	34.1	46.1	25.5
+ **NCA**	60.8 ↓	**3.3**	26.8 ↓	**32.3**	**11.7**	**11.0**	**65.3**	**74.3**	**35.7**
Mixtral-8×7B-SFT	**75.6**	16.7	61.0	57.6	40.1	25.9	85.9	87.5	56.3
+ DPO	74.9 ↓	17.2	47.6 ↓	55.8 ↓	35.3 ↓	**26.9**	67.3 ↓	75.7 ↓	50.1↓
+ **NCA**	**75.6**	**21.1**	**62.8**	**61.5**	**41.6**	**26.9**	**86.8**	**86.9**	**57.9**

4) NCA effectively prevents chosen likelihood from decreasing.