Rebuttal Supplements

Table 1. Model evaluation with beaver-7b-unified-reward/cost models on PKU-SafeRLHF test set.

Initial Model	Optm.	Δ Helpfulness $\uparrow$	Harmlessness $\downarrow$	S.R.
Alpaca-7B	Initial	-	1.59	40.83%
	SafeRLHF	-0.36	-3.13	89.63%
	SACPO	-0.16	-2.22	83.94%
	RePO	+1.11	-4.31	96.14%
Llama3.2-3B	SFT	-	1.92	36.92%
	SafeRLHF	+0.13	-2.11	77.18%
	SACPO	-1.46	0.79	52.78%
	RePO	-0.95	-2.12	82.93%

Table 2. GPT-4 evaluation of compared initial model and various optimization algorithms on the whole PKU-SafeRLHF test set.

Initial v.s.	RePO	SafeRLHF	SACPO
Alpaca-7B	(10.24%, 81.67%)	(11.76%, 70.73%)	(26.99%, 61.25%)
Llama3.2-3B	(18.77%, 72.88%)	(16.06%, 64.92%)	(48.99%, 28.51%)

Table 3. GPT-4 evaluation of compared RePO and other baselines on the whole PKU-SafeRLHF test set.

Initial	RePO v.s. SafeRLHF	RePO v.s. SACPO
Alpaca-7B	(52.59%, 21.43%)	(76.99%, 13.34%)
Llama3.2-3B	(37.10%, 45.32%)	(87.04%, 9.36%)

Table 4. GPT-4 evaluation of safety rate on the whole PKU-SafeRLHF test set.

	Initial	RePO	SafeRLHF	SACPO
Alpaca-7B	44.75%	96.21%	85.90%	75.60%
Llama3.2-3B	41.97%	85.65%	75.98%	36.03%

Table 5. XSTest results.

Initial Model	Alpaca-7B		Llama3.2-3B	
Optim.	RePO	SafeRLHF	RePO	SafeRLHF
Over Refusal ↓	20.40%	6.80%	14.00%	8.0 %
Success Refusal ↑	72.00%	46.50%	68.00%	39.5%

Table 6. OrBench results.

Initial Model	Alpaca-7B		Llama3.2-3B	
Optim.	RePO	SafeRLHF	RePO	SafeRLHF
seemingly toxic prompts rejection rate ↓	57.39%	33.66%	43.59%	21.07 %
toxic prompts rejection rate ↑	82.60%	70.84%	71.76%	36.79%

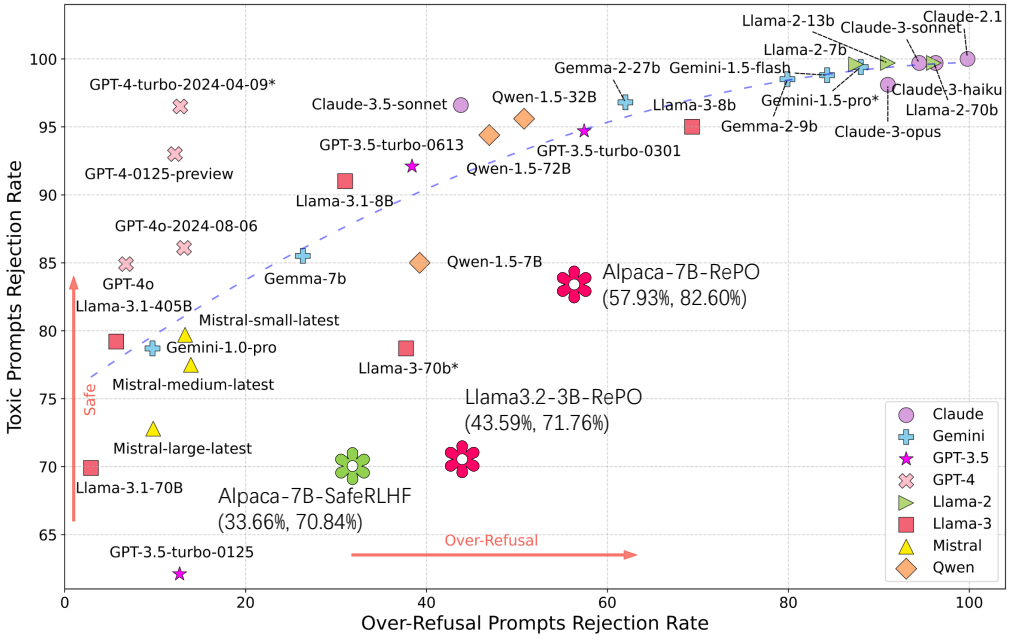

Figure 1. The schematic positions of RePO and SafeRLHF on OrBench. The original figure is Figure 1 of Orbench. In the figure, the slope of the blue line is determined by the quadratic regression coefficient of all the points to represent the overall performance of all models. On this basis, we have roughly marked the positions of the four models shown in Table 6 in the figure. The red flowers represent the Alpaca-7B and Llama3.2-3B optimized by RePO, and the green flower represents the Alpaca-7B optimized by SafeRLHF. Since the safety of the Llama3.2-3B optimized by SafeRLHF is too low and exceeds the lower limit of the original figure, we did not mark it in the figure.

Table 7. The GPT-4 Evaluation results on BeaverTails-30K test set (The size is 2760 after removing the duplicate prompts) of the serious models whose initial models are Alpaca-7B. The evaluation prompts are the same as Table 2 and Table 3 in the submission. The first line compares RePO with the initial model and other optimization algorithms, the left is RePO's win rate and the right is other models' win rate. The second line is the safety rate of each model.

	Initial	SafeRLHF	SACPO	RePO
RePO v.s.	(81.66%, 11.56%)	(45.93%, 27.10%)	(74.96%, 15.25%)	-
S.R.	40.25%	82.61%	69.64%	93.62%