Included here are additional evaluation results (highlighted in orange) on RealToxicityPrompts-Gen, IMDB-Gen and Jigsaw-Gen over GPT-2 base model. We now include evaluation over text diversity measuring metric Distinct-n (Dist-n) introduced in DExperts [1]. Dist-n measures diversity as unique n-grams count normalized by the text length. We add three more safety inducing baselines: 1) GPT-2 in-context prompted to generate safe text. We add the prompts 'Generate positive sentiment' and 'Generate non-toxic text' for respective datasets. 2) DExperts [1]: A test-time decoding method. 3) Quark [2]: A finetuning based method. We could not add results on RECT [3] as their model checkpoints are not public. For Dexperts, as suggested in the work, we used GPT-2 as the expert and the author provided GPT-2 anti-expert checkpoint. For Quark, we used the finetuned toxicity free GPT-2 Large (762M parameters) model to obtain generations on RealToxicityPrompts-Gen and Jigsaw-Gen. We used their GPT-2 Large sentiment steering model to obtain generations on IMDB-Gen. For toxicity datasets, the toxicity score is obtained using `unitary/toxic-bert` LLM that outputs how toxic a piece of text is. To obtain sentiment scores, we use `lvwerra/distilbert-imdb` LLM. We evaluate all the algorithms on worst scoring prompts from the tail of score distribution over 5k randomly sampled prompts from the test datasets. The cutoff point for tail are scores of: -2.5 (IMDB-Gen), -5 (RealToxicityPrompts-Gen) and -5 (Jigsaw-Gen). We observe that across datasets, models returned by our proposed algorithm RA-RLHF enjoy the best performance in terms of safety scores while maintaining text coherence and diversity.

1 Extended results on IMDB-Gen

Model	Sentiment Score (↑)	Dist-1	Dist-2	Dist-3
GPT-2	-2.607	0.902	0.969	0.946
Prompted	-2.595	0.910	0.960	0.935
DExperts	-2.635	0.933	0.897	0.824
SFT	-2.465	0.916	0.963	0.937
RLHF	-1.299	0.883	0.965	0.945
Quark	-2.008	0.833	0.952	0.940
RA-RLHF	-0.908	0.874	0.967	0.948

Table 1: Sentiment score and diversity evaluation metrics

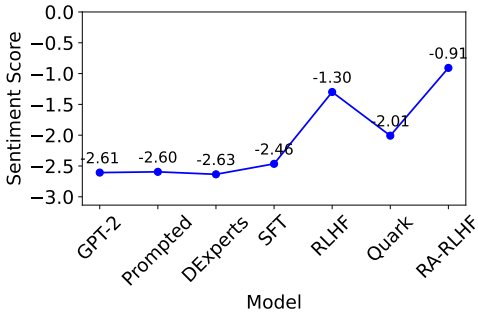

Figure 1: Tail score plotted

2 Extended results on RealToxicityPrompts-Gen

Model	-ve Toxicity Score (↑)	Dist-1	Dist-2	Dist-3
GPT-2	1.662	0.937	0.952	0.911
Prompted	1.663	0.937	0.949	0.906
DExperts	1.587	0.932	0.883	0.809
SFT	1.162	0.919	0.954	0.916
RLHF	2.426	0.912	0.956	0.921
Quark	2.587	0.883	0.945	0.913
RA-RLHF	2.834	0.904	0.956	0.922

Table 2: Negative toxicity score and diversity evaluation metrics

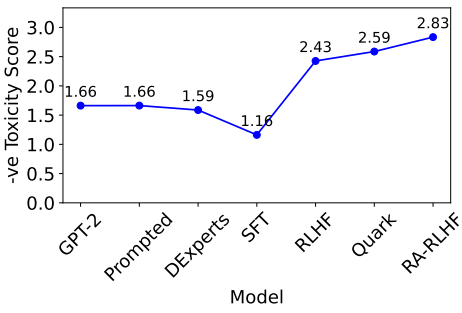

Figure 2: Tail score plotted

3 Extended results on Jigsaw-Gen

Model	-ve Toxicity Score (↑)	Dist-1	Dist-2	Dist-3
GPT-2	0.348	0.933	0.933	0.886
Prompted	0.614	0.945	0.942	0.893
DExperts	0.422	0.883	0.852	0.792
SFT	0.532	0.938	0.945	0.900
RLHF	1.694	0.916	0.940	0.902
Quark	1.521	0.870	0.920	0.885
RA-RLHF	2.057	0.913	0.956	0.923

Table 3: Negative toxicity score and diversity evaluation metrics

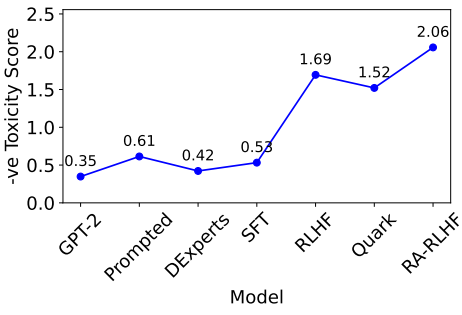

Figure 3: Tail score plotted