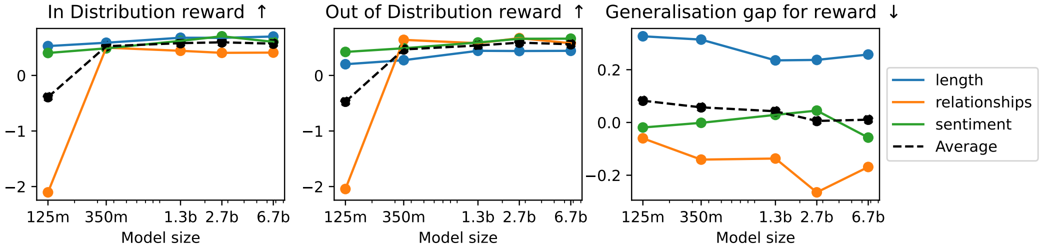

Figure 21: Proxy RM Score for RL models for each dataset split, both in-distribution and out-of-distribution performance, and the generalisation gap. Arrows ↑, ↓ indicate whether higher or lower scores are better.

Table 11: Per-input diversity scores for both RLHF and SFT models averaged over dataset splits. For these scores the outputs used to calculate the diversity are a sample of outputs from the model for single input. These per-input scores are then averaged, as in Eq. (2). Bolded results are better scores for each model size.

Model Size	125m		350m		1.3b		2.7b		6.7b	
Model Type	RLHF	SFT	RLHF	SFT	RLHF	SFT	RLHF	SFT	RLHF	SFT
EAD	0.15	0.81	0.15	0.8	0.13	0.81	0.16	0.8	0.07	0.79
Sent BERT	0.15	0.5	0.12	0.46	0.15	0.48	0.13	0.45	0.06	0.45
NLI	-1.08	0.26	-0.96	0.2	-1.0	0.12	-1.07	0.09	-1.54	0.06
Average	-0.26	**0.52**	-0.23	**0.49**	-0.24	**0.47**	-0.26	**0.45**	-0.47	**0.44**

Table 12: Across-input diversity scores for both RLHF and SFT models averaged over dataset splits. For these scores the outputs used to calculate the diversity are a set of single outputs from a range of inputs, as in Eq. (3). Bolded results are better scores for each model size.

Model Size	125m			350m			1.3b			2.7b			6.7b		
Model Type	RLHF	SFT	BoN	RLHF	SFT	BoN	RLHF	SFT	BoN	RLHF	SFT	BoN	RLHF	SFT	BoN
EAD	0.69	0.86	0.85	0.87	0.87	0.86	0.86	0.87	0.86	0.86	0.87	0.86	0.87	0.87	0.86
Sent BERT	0.63	0.73	0.72	0.71	0.73	0.71	0.68	0.73	0.71	0.7	0.74	0.71	0.73	0.73	0.71
NLI	0.05	0.35	0.32	0.2	0.38	0.3	0.25	0.36	0.27	0.3	0.28	0.27	0.32	0.32	0.3
Average	0.46	**0.64**	0.63	0.59	**0.66**	0.62	0.6	**0.65**	0.61	0.62	0.63	0.62	0.64	0.64	0.62

Table 13: Across-input diversity scores for BoN models averaged over dataset splits. For these scores the outputs used to calculate the diversity are a set of single outputs from a range of inputs, as in Eq. (3). This is equivalent to Table 2 in the main paper but for BoN policies. See response to reviewer hdr6 for more details.

Model Size	125m	350m	1.3b	2.7b	6.7b
EAD	0.85	0.86	0.86	0.86	0.86
Sent BERT	0.72	0.71	0.71	0.71	0.71
NLI	0.32	0.3	0.27	0.27	0.3
Average	0.63	0.62	0.61	0.62	0.62

Table 14: SFT Model ROUGE1 and reward model accuracy for the 3 smallest OPT model sizes, comparing freezing 80% of layers vs no layer freezing. Freezing generally results in less performance, as expected.

Model Size	SFT Model Rouge1		RM Accuracy	
	80% Frozen	0% Frozen	80% Frozen	0% Frozen
125m	0.217	0.221	0.482	0.496
350m	0.2241	0.2233	0.498	0.508
1.3b	0.221	0.2347	0.538	0.559

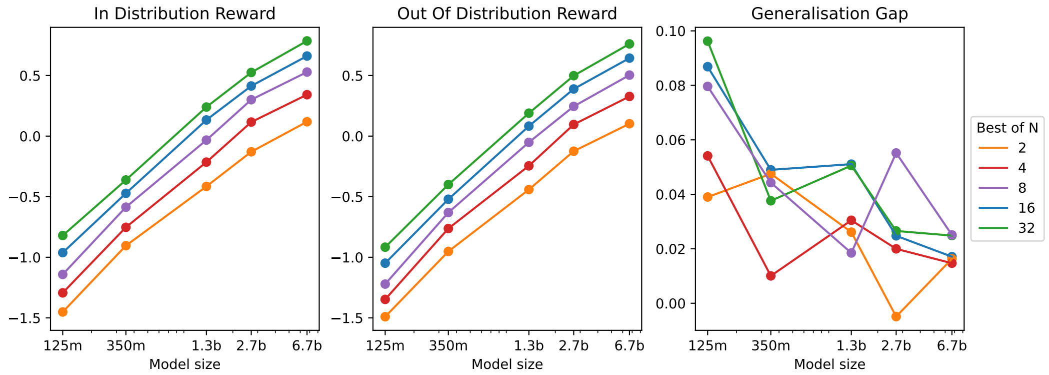

Figure 22: Proxy RM score for BoN sampling with varying N. All metrics are averaged over the 3 dataset splits.

for models with more than 125 million parameters both SFT and RLHF produce policies with very similar scores.

However, for the Sent BERT score, which is a proxy for diverse content and semantics, RLHF models are consistently less diverse than SFT models. This implies that RLHF produces models that have a tendency to generate outputs about certain topics or content regardless of the input. For the NLI score, which is a proxy for logical diversity, we see that as RLHF model size increases the score increases, eventually reaching the diversity of SFT models. Low NLI score for smaller models implies they have a tendency to make logically consistent claims in their outputs on top of producing outputs about certain topics of content.

J.5 Model Freezing Experiments

We perform a small experiment to evaluate the effects of freezing the first 80% of layers during fine-tuning. The results are shown in Table 14, and show that while performance drops for the models evaluated in the experiment, the drop is not catastrophic, justifying the use of model freezing.

J.5.1 BoN Performance for Different N

In the previous OPT results we use $N = 64$ samples in best of N sampling. Here we show proxy RM scores for $N = 2, 4, 8, 16, 32$, to show how choices of N trade off against performance. Fig. 22 shows proxy RM score in-distribution, out-of-distribution and the generalisation gap for these different choices of N. We see that increasing N does lead to improved performance. At lower model sizes it leads to a larger generalisation gap, but this does not hold for larger model sizes.

K RLHF and RM Training Curves

Here we present training curves for PPO and reward model training, for the summarisation task, for the models used in the main paper. In Fig. 23 we show the reward model validation accuracy throughout training. This is 1 epoch of training. In Fig. 24 we show the KL divergence and reward model score throughout training. PPO training has converged by approximately 250 PPO training steps, and so we terminated training early to save compute.

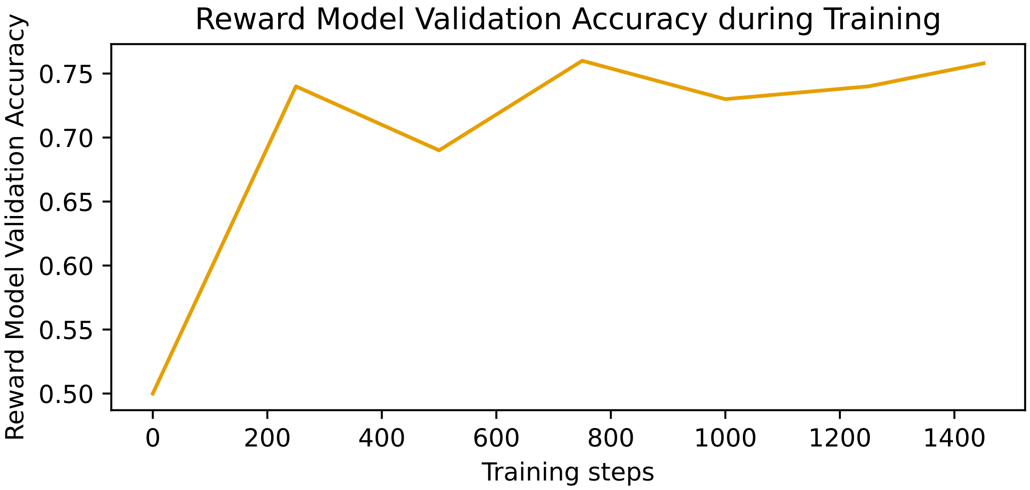

Figure 23: Reward model validation accuracy during training for the summarisation task.

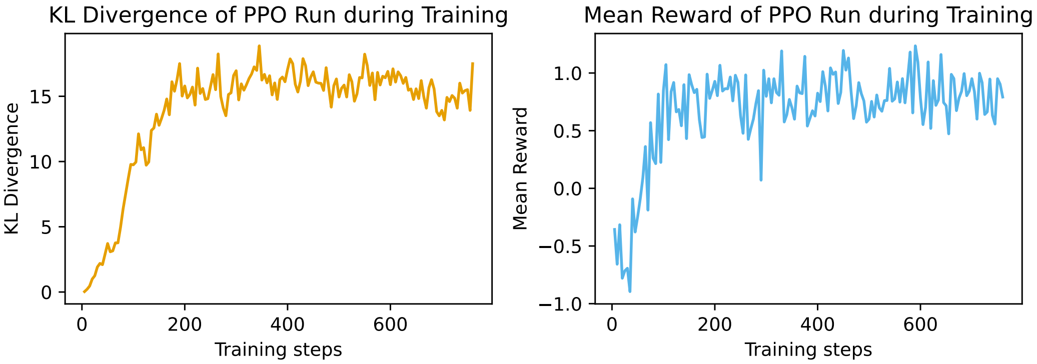

Figure 24: PPO KL divergence and reward model score curves for the summarisation task.