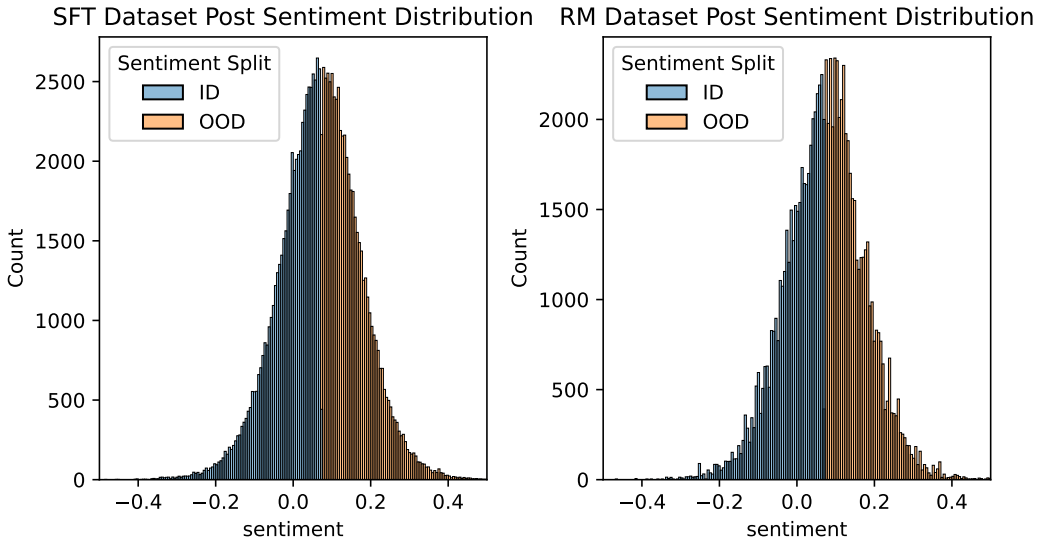

Figure 15: The distribution of post sentiment across the full SFT and RM datasets. ID is the in-distribution version of the dataset, and OOD is the out-of-distribution version.

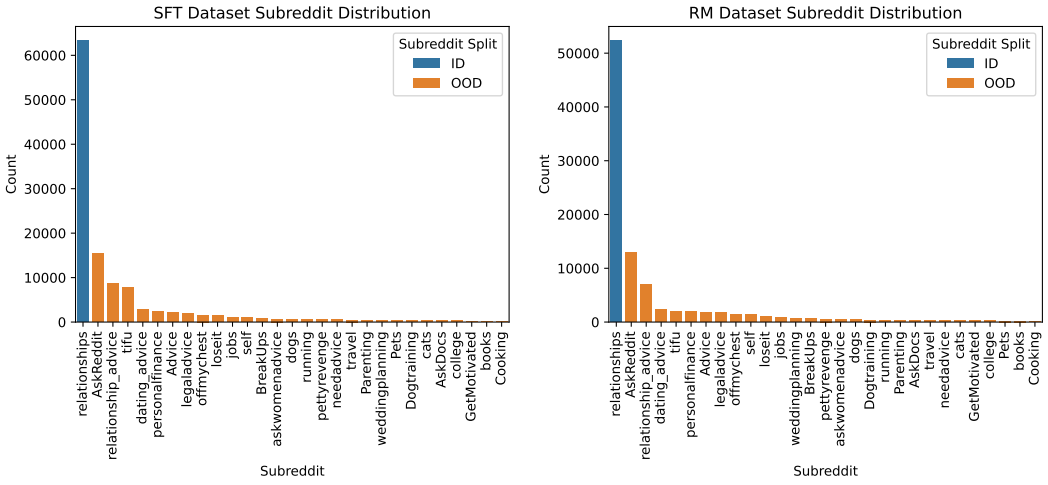

Figure 16: The number of posts in each subreddit across the full SFT and RM datasets. ID is the in-distribution version of the dataset, and OOD is the out-of-distribution version.

Table 9: Learning rates for different model sizes and dataset splits for reward models. Underlined learning rate is the chosen one. ke^{-n} means $k \times 10^{-n}$.

Dataset	125m	350m	1.3b	2.7b	6.7b
relationships	$\underline{5e^{-4}}$,$1.5e^{-5}$,$5e^{-5}$	$\underline{5e^{-4}}$,$1.5e^{-5}$,$5e^{-5}$	$\underline{5e^{-5}}$,$1.5e^{-5}$,$5e^{-6}$	$5e^{-5}$,$\underline{1.5e^{-5}}$,$5e^{-6}$	$\underline{5e^{-5}}$,$1.5e^{-5}$,$5e^{-6}$
length	$\underline{5e^{-4}}$,$1.5e^{-5}$,$5e^{-5}$	$5e^{-4}$,$1.5e^{-5}$,$\underline{5e^{-5}}$	$5e^{-5}$,$\underline{1.5e^{-5}}$,$5e^{-6}$	$5e^{-5}$,$\underline{1.5e^{-5}}$,$5e^{-6}$	$5e^{-5}$,$1.5e^{-5}$,$\underline{5e^{-6}}$
sentiment	$5e^{-4}$,$\underline{1.5e^{-5}}$,$5e^{-5}$	$5e^{-4}$,$1.5e^{-5}$,$\underline{5e^{-5}}$	$\underline{5e^{-5}}$,$1.5e^{-5}$,$5e^{-6}$	$5e^{-5}$,$\underline{1.5e^{-5}}$,$5e^{-6}$	$\underline{5e^{-5}}$,$1.5e^{-5}$,$5e^{-6}$

Table 10: Learning rates for different model sizes and dataset splits for RLHF models. Underlined learning rate is the chosen one. ke^{-n} means $k \times 10^{-n}$.

Dataset	125m	350m	1.3b	2.7b	6.7b
sentiment	$1e^{-4}$,$3e^{-5}$,$\underline{1e^{-5}}$	$1e^{-4}$,$3e^{-5}$,$\underline{1e^{-5}}$	$\underline{1e^{-5}}$,$3e^{-6}$,$1e^{-6}$	$\underline{3e^{-5}}$	$\underline{1e^{-5}}$
length	$1e^{-4}$,$3e^{-5}$,$\underline{1e^{-5}}$	$1e^{-4}$,$3e^{-5}$,$\underline{1e^{-5}}$	$\underline{1e^{-5}}$,$3e^{-6}$,$1e^{-6}$	$\underline{3e^{-5}}$	$\underline{5e^{-6}}$
relationships	$1e^{-4}$,$3e^{-5}$,$\underline{1e^{-5}}$	$1e^{-4}$,$3e^{-5}$,$\underline{1e^{-5}}$	$\underline{1e^{-5}}$,$3e^{-6}$,$5e^{-6}$	$\underline{5e^{-6}}$,$3e^{-6}$,$1.7e^{-6}$	$\underline{5e^{-6}}$,$3e^{-6}$,$1.7e^{-6}$

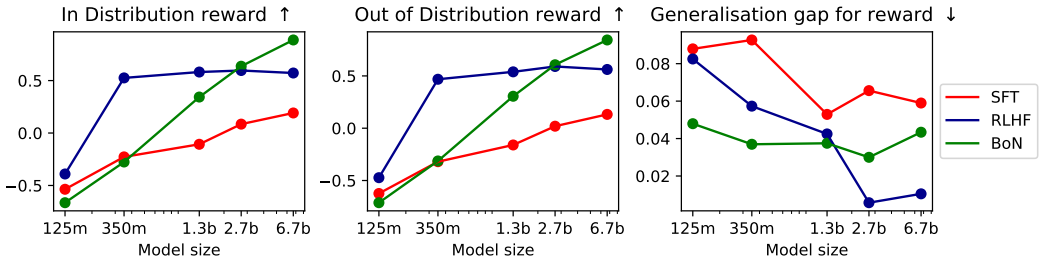

Figure 17: Proxy RM Score for SFT, BoN and RL models, averaged over dataset splits, for both in-distribution and out-of-distribution performance, and the generalisation gap. Arrows ↑, ↓ indicate whether higher or lower scores are better.

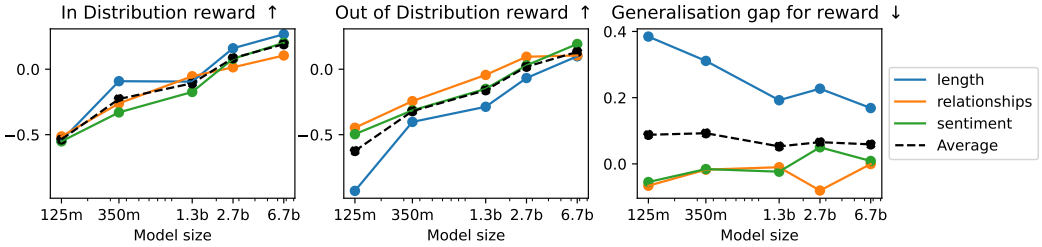

Figure 18: Proxy RM Score for SFT models for each dataset split, both in-distribution and out-of-distribution performance, and the generalisation gap. Arrows ↑, ↓ indicate whether higher or lower scores are better.

J.3 Generalisation Evaluation

For evaluation in these experiments, we train an RM as described in Appendix E.1 using the full dataset of summaries and preferences without splitting and the 6.7 billion parameter OPT model. We then use this *proxy RM* to evaluate the performance of SFT, BoN and RLHF models. As this reward model is trained with a different random seed and a different data distribution, it serves as a held-out automated evaluation of models, and a good proxy for human preferences.

We first discuss the results from the OPT version of the generalisation experiments described in Section 5.1. Fig. 17 shows the proxy RM score for SFT, BoN and RLHF models, averaged over dataset sizes. The important result here is that BoN generalises better than RLHF, which generalises better than SFT.

We see that at middling model sizes, RLHF outperforms BoN, but BoN scales better than RLHF, eventually outperforming it for models with more than 2.7b parameters. SFT comes out worst in this comparison, both scaling worse than BoN and the same as RLHF, and having worse absolute performance and generalisation. We see that BoN sampling does not see diminishing returns as model size increases, implying it will continue to be a useful yet simple technique. This experiment highlights the importance of training a reward model on human feedback and using it to select the best outputs at test time, potentially after fine-tuning the model.

Supervised Fine-Tuning. Fig. 18 shows the proxy RM score for the SFT models. Performance increases as model size increases, and in general performance drops OOD, which is unsurprising. There is a slight downward trend in the average generalisation gap across dataset splits as model size increases, implying that larger models fine-tuned with SFT generalise better (in terms of automated metrics). The relationships and sentiment splits produce negligible generalisation gap for the proxy RM score while the length split is more difficult.

Best-of-N. Fig. 19 shows the proxy reward score for the BoN models. Here we can see a smooth almost linear increase in performance as the model size increases. Given that the RM scores for smaller models were below chance, the fact the even for smaller models BoN results improve performance requires explanation. We hypothesise that the smooth increase here comes from two factors: the improvement in the SFT model being sampled from, and the improvement from the

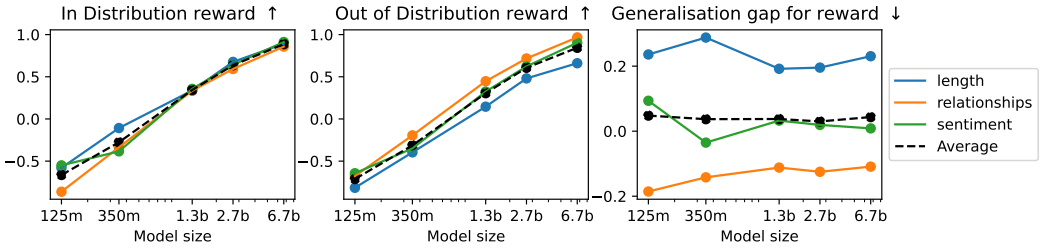

Figure 19: Proxy RM Score for BoN models for each dataset split, both in-distribution and out-of-distribution performance, and the generalisation gap. Arrows ↑, ↓ indicate whether higher or lower scores are better.

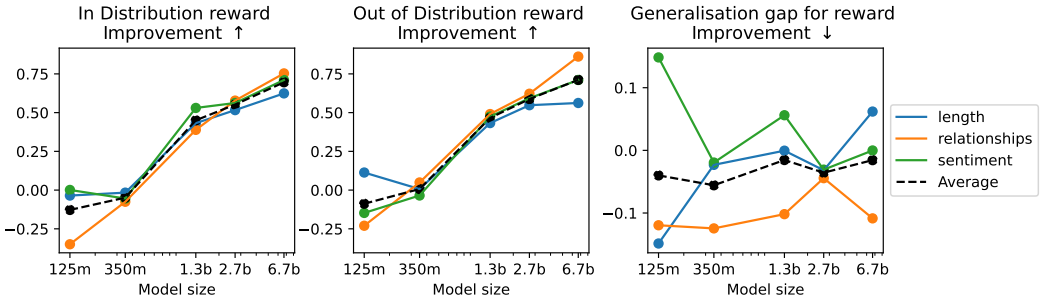

Figure 20: Proxy RM Score improvement using BoN on top of SFT models for each dataset split, both in-distribution and out-of-distribution performance, and the generalisation gap. Arrows ↑, ↓ indicate whether higher or lower scores are better. This plot is highlights the improvement from Fig. 18 to Fig. 19.

RM. At smaller model sizes increasing the number of parameters leads to improvements in SFT performance but not RM performance, while at larger model sizes increasing the number of parameters leads to improvements in both SFT and RM performance, but both to a lesser extent.

Fig. 20 shows the improvement of BoN over SFT. We can see that BoN only starts improving SFT as model size passes 350 million parameters, and the improvement grows as model size grows. This implies that as we increase model size BoN is likely to become a more performant choice compared to SFT. Further, we note that BoN uniformly improves the generalisation across model sizes (as shown by the black dashed line), implying that scaling models, even using SFT+BoN, will still result in a non-zero generalisation gap.

Reinforcement Learning from Human Feedback. Fig. 21 shows the proxy RM score for the RLHF models. Again, we see that increasing model size improves performance. Here we see a clearer trend of increasing model sizes reducing generalisation gap – this implies that as we make RL models larger, they are likely to generalise better. Given the difference between this trend and the generalisation gap trend for SFT models in Fig. 18, this partially justifies why RLHF is used in fine-tuning LLMs at a very large scale (Glaese et al., 2022; OpenAI, 2023; Ouyang et al., 2022): RLHF produces better-generalising models at larger model sizes than SFT.

J.4 Diversity Evaluation

Table 11 shows the per-input diversity scores (Eq. (2)) for both RLHF and SFT models. The RLHF models have much lower diversity than SFT models according to all three metrics. While RLHF leads to better-generalising policies, those policies generate much less diverse outputs. We also see that diversity does not seem to change much with model size, apart from a slight downward trend for diversity in SFT models as model size increases.

Table 12 shows the across-input diversity scores (Eq. (3)). Here the corpus of text over which the diversity is measured is a single input sampled from the model for a range of outputs. Even if a model produces less diverse outputs for a single input, it could still produce different inputs for different outputs. This is the case for the EAD score, which is a proxy for diverse vocabulary and syntax, as