

Figure 2: **Summarisation Generalisation Results.** GPT-4-PvR for SFT, BoN and RL policies, based on LLaMa 7B, trained on the summarisation task. In-distribution is performance on TL;DR, and out-of-distribution is on CNN/DailyMail, and generalisation gap is ID – OOD performance.

generalisation gap; Bo16 outperforms RLHF which outperforms SFT, both ID and OOD. While Bo16 outperforming RLHF is somewhat surprising, there are examples of this in the literature (Nakano et al., 2022), and Bo16 has a substantially higher inference-time cost, making RLHF better for practical applications. The performance of all policies drops OOD, which is unsurprising given the difficulty of the shift (reddit post summarisation to news article summarisation). The generalisation gap is fairly similar between methods, implying that none of these methods has a particular advantage with respect to generalisation specifically in this setting.

These results also show that the generalisation gap for SFT and Bo16 policies are the same (and they are still the same for a different choice of temperature for Bo16, see Fig. 9). This can be explained by the fact that the reward model generalises near-perfectly to the CNNDM preference dataset from (Stiennon et al., 2022). ID and OOD accuracy is 75.8 and 71.6 respectively, and considering that the maximum inter-annotator agreement in the CNNDM preference dataset is lower than TL;DR (and hence the maximum accuracy attainable is lower), it is plausible that the RM is not suffering any real drop in OOD performance in this case. This implies that all of the drop in performance for Bo16 is driven by the drop in performance of the SFT model, as if both SFT and RM performed worse OOD we would expect those drops in performance to compound. Overall, this shows that if you can expect your reward model to generalise well then BoN is a good choice of policy, although it is limited by the generalisation abilities of the underlying model being sampled from (the SFT model in this case), and is more expensive at inference time than RLHF and SFT.

In Appendix J we present results for a range of models based on OPT (Zhang et al., 2022), trained in a similar way to the LLaMa models but on split versions of the TL;DR dataset, and evaluated with a LLaMa proxy reward model. These results show similar trends: BoN outperforms RLHF which in turn outperforms SFT at the largest model size, and the ordering holds OOD for 3 different splits of the training dataset. This shows our results are robust across different evaluation metrics and base models.

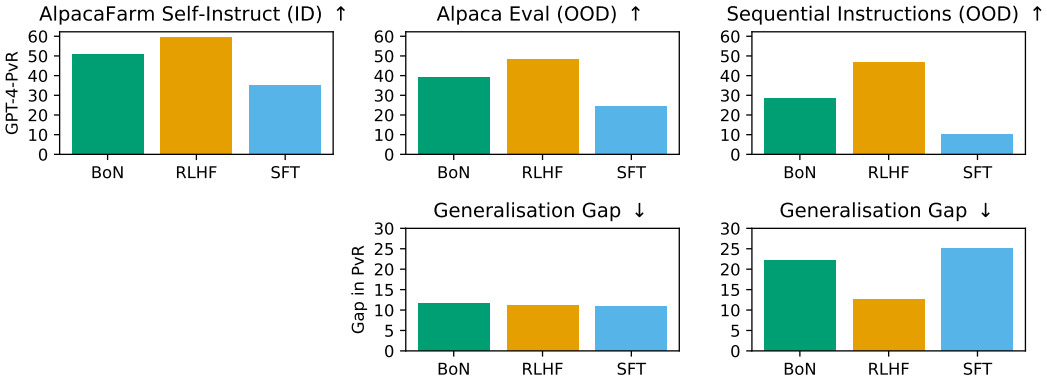


Figure 3: **Instruction Following Generalisation Results.** GPT-4 PvR for SFT, BoN and RL policies, based on LLaMa 7B, trained on the AlpacaFarm Self-Instruct instruction following task. ID is on AlpacaFarm Self-Instruct, OOD is on the AlpacaEval and Sequential Instructions datasets respectively, and generalisation gap is ID – OOD performance.

Instruction Following. Fig. 3 shows the results of BoN, SFT and RLHF models trained in AlpacaFarm Self-Instruct (Dubois et al., 2023) evaluated ID and OOD on AlpacaEval and Sequential Instructions. Similar to summarisation, we see that RLHF and Bo16 both outperform SFT, but here RLHF outperforms Bo16 across all datasets, in contrast to the summarisation task. As the focus of this paper is mostly comparing RLHF and SFT, we did not investigate this result further, but there could be many possible reasons for the change in ordering between RLHF and Bo16, as the two tasks are very different and the model training procedures are not identical.

We see that on AlpacaEval (the easier OOD generalisation task), models all generalise equally well, but on the harder Sequential Instructions OOD task, RLHF generalises much better. This suggests that *RLHF may generalise better relative to SFT for larger distribution shifts*, which potentially explains why models fine-tuned with RLHF have been observed to be much better in practice when interacting with users (Touvron et al., 2023b; Ouyang et al., 2022, *inter alia*): when users interact with these models the distribution shift is quite pronounced and hence many inputs are more OOD, and this is where RLHF model performance continues to be high.

While GPT-4 PvR is a useful metric, it does not show a difference in generalisation (as measured by generalisation gap) between SFT, RLHF and Bo16 models on the easier AlpacaEval dataset. This could be due to these models having similar generalisation properties, or be a deficiency of the metric. To investigate this further, we use can look at the GPT-4 Head-to-Head winrate of SFT vs Bo16 and vs RLHF, which is shown in Fig. 4. These results show that both RLHF and Bo16 winrates *improves* vs SFT by approximately 3.5% from ID to AlpacaEval OOD. This implies that RLHF and Bo16 generalise better than SFT even in this case, emphasising the need for a range of metrics when evaluating model generalisation.

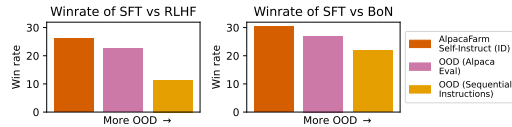


Figure 4: **GPT-4 Head-to-Head Winrate** of SFT vs RLHF and Bo16 in AlpacaFarm Self-Instruct, AlpacaEval and Sequential Instructions datasets.

6.2 DIVERSITY

For the diversity evaluations, we focus on the summarisation task specifically, as it has the most compelling results. We ran some initial experiments evaluating diversity for the instruction-following models, but we did not see any meaningful differences. We hypothesise this is due to the diversity metrics we use being designed for settings where the model output is relatively short (e.g. a single sentence), whereas in the instruction-following setting outputs are generally much longer. Furthermore, RLHF models tend to produce longer outputs than other models, which can confound the evaluation of output diversity, since most metrics are not invariant to output length.

Fig. 5 shows the per-input diversity scores for RLHF and SFT models in the summarisation task. We see that across the first two metrics, RLHF has much lower output diversity than SFT. Fig. 6 shows the across-input diversity scores in the same setting. Here we see that while SFT generally has slightly higher diversity as before, the difference is much smaller than in the per-input case. The drop in across-input diversity cannot be explained purely by the use of the reward model, as BoN has similar or higher across-input diversity than SFT for the first two metrics. Both these trends are the same for OPT across model sizes (see Appendix J.4 for full results) We find that NLI does not show meaningful difference between models in either per-input or across input, showing that in a logical sense all models are similarly diverse.

The difference in across-input diversity between RLHF and SFT, while small, can also be taken as evidence of the phenomena of “mode collapse” hypothesised to occur under RLHF fine-tuning (janus, 2022). The hypothesised effect is that even for different inputs, RLHF models can be biased towards outputting text of a specific style or “mode”, meaning that even changing the inputs to a model is not sufficient to generate truly diverse outputs. We believe that this is the first rigorous empirical demonstration of across-input mode collapse emerging from RLHF training specifically.

For most metrics and models, the across-input diversity scores are higher than the per-input diversity scores, which is expected given the across-input diversity distribution is much broader. However, for EAD (which measures diversity at the n-gram level), SFT has similar levels of per-input and

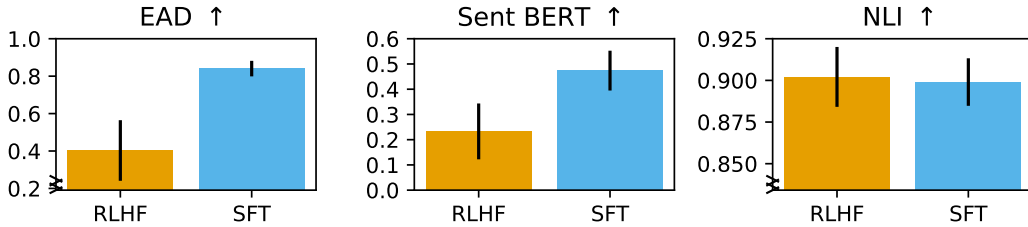


Figure 5: **Per-input diversity metrics for RLHF and SFT models.** For these scores the outputs used to calculate the diversity are a sample of outputs from the model for single input. These per-input scores are then averaged, as in Eq. (2). Error bars are standard deviation of the per-input diversity score across different inputs. Note that some plots have broken y-axis for better visualisation.

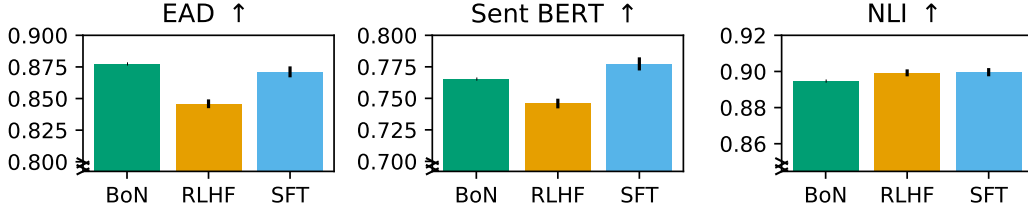


Figure 6: **Across-input diversity metrics for RLHF, BoN and SFT models.** For these scores the outputs used to calculate the diversity are a set of single outputs from a range of inputs, as in Eq. (3). Note that all plots have broken y-axis for better visualisation; the differences between SFT and RLHF are much smaller in this case than in the per-input diversity metrics in Fig. 5. Error bars (where present) are standard deviation of the across-input scores over different samples from the set of outputs for each input.

across-input diversity. This is likely due to SFT effectively approaching the maximum EAD diversity even in the per-input case, so that the across-input diversity cannot be much higher.

6.3 THE IMPACT OF THE KL PENALTY

We have shown that while RLHF improves performance ID and OOD in an absolute sense, this comes at the cost of substantial drops in output diversity relative to SFT. Motivated by the fact that the KL penalty coefficient (see Eq. (1)) encourages the RLHF policy to stay closer to the SFT policy, we investigate whether adjusting this coefficient trades off between generalisation and diversity. The results show that this does not work – *increasing the KL penalty coefficient leads to a drop in performance as expected, but also to a drop in per-input diversity*, rather than a gain (see Appendix I for details). This emphasises that more research is needed to investigate whether more sophisticated methods can improve the trade-off between generalisation and diversity.

7 DISCUSSION AND CONCLUSION

Summary of Contributions. In this work, we analyse three methods for fine-tuning LLMs (RLHF, SFT, and BoN) across two problem settings (summarisation and instruction following) in terms of OOD generalisation and output diversity. We demonstrate an inherent tradeoff between generalisation performance and output diversity when choosing between these methods: RLHF produces more performant models both in-distribution and out-of-distribution, but at the cost of lower output diversity, both per-input and across-input. It is unclear whether this tradeoff is a fundamental one in fine-tuning LLMs with RLHF or just demonstrates a deficiency in current methods. We suspect the answer will be a combination of both explanations: There will be a pareto-frontier of output diversity vs generalisation performance on which tradeoffs have to be made, but current methods do not yet seem to be at that frontier. Future work could investigate producing methods that are closer to this frontier, either through increasing the performance of SFT or increasing the output diversity of RLHF.