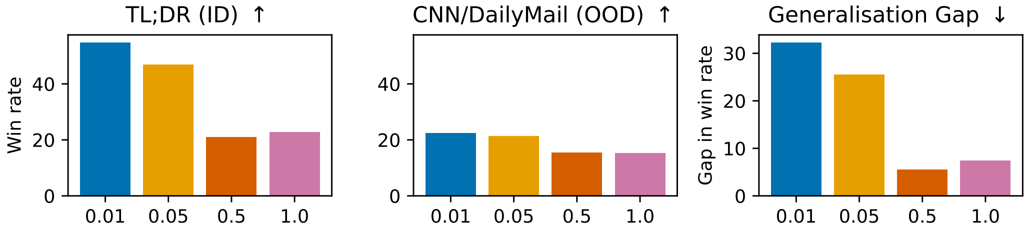

Figure 11: GPT-4 API evaluation win rate vs reference (text-davinci-003) outputs for RLHF models, based on LLaMa 7B, trained on the summarisation task, sweeping over KL penalty coefficient.In-distribution is performance on TL;DR, out-of-distribution is on CNN/DailyMail, and generalisation gap is ID – OOD performance.

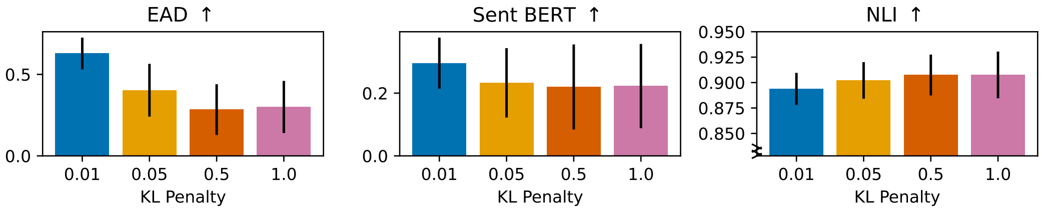

Figure 12: Per-input diversity metrics for RLHF summarisation models with different KL penalty coefficients. For these scores the outputs used to calculate the diversity are a sample of outputs from the model for single input. These per-input scores are then averaged, as in Eq. (2). Error bars are standard deviation of the per-input diversity score across different inputs. Note that some plots have broken y-axis for better visualisation.

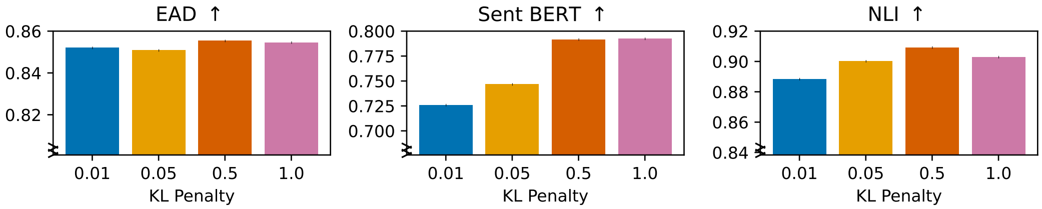

Figure 13: Across-input diversity metrics for RLHF, Bo16 and SFT policies. For these scores the outputs used to calculate the diversity are a set of single outputs from a range of inputs, as in Eq. (3). Note that all plots have broken y-axis for better visualisation.

Table 6: Size of the train, validation, test and OOD test datasets for each split for the SFT and RLHF models.

	length	sentiment	relationships
Train	58770	58361	63324
Validation	3234	3223	3462
Test	3303	3276	3539
OOD Test	3250	3277	3014

J Summarisation Experiments with OPT

In addition to the experiments in the main paper, we did several experiments using OPT (Zhang et al., 2022) in the summarisation task with a different choice of ID and OOD test sets.

J.1 Dataset Splitting

We create split versions of these datasets along several factors of variation in their inputs: *length*, *sentiment*, and *subreddit*. For each of these factors of variation, we create a train/test split where the train and test inputs are drawn from different parts of the distribution for that factor. For the *length* split, the training set consists of data where the post is less than 245 words (the median number of words in the SFT training distribution); for the *sentiment* split, we use an off-the-shelf sentiment classifier (Loria, 2013), and the training set consists of summaries with sentiment less than the median sentiment in the dataset; for the *subreddit* split, the training set consists of summaries from a specific subreddit, r/relationships. In all cases the test set is the complement of the training set in the full dataset, meaning that the trained models will be evaluated on inputs from a different distribution than the one seen during training.

In all cases, we apply the same splitting procedure to both the preference data and the input/output pairs, to ensure that the training and test sets are consistent across the different methods. Each of these splits was chosen to produce a roughly 50-50 split between the training and testing distributions. In the case of *length* and *sentiment* this is exact, and in the case of *subreddit* the r/relationships subreddit contains approximately 60% of the data.

While we do not expect these splits to capture the full range of distribution shifts models may experience when deployed, using a range of splits will give us a more robust measure of how well the policies trained with different methods generalise under distribution shifts.

The dataset we use from (Stiennon et al., 2022) (filtered from (Nallapati et al., 2016)) comes with train, validation and test splits, which we use throughout our work. For the SFT dataset these splits have size 116722, 6553 and 6447 respectively, and for the RM dataset they have size 92858, 33083, 50718 respectively. The SFT and RLHF splits are the same apart from the RLHF dataset does not require the summaries (outputs), just the posts (inputs). To create the dataset splits used for the OOD generalisation experiments in Section 6.1, we split each of the train, validation and test sets into an in-distribution (ID) and out-of-distribution (OOD) train, validation and test set. We then train on the ID train set, do model selection using the ID validation set, and evaluate on the ID and OOD test sets to measure the in-distribution and out-of-distribution performance.

For the sentiment split, we first measure the sentiment of each post using an off-the-shelf sentiment classifier (Loria, 2013). For a given subset of the dataset (i.e. train, validation, test), the ID version of that subset is the set of all inputs with posts whose sentiment is lower than the median sentiment, and the OOD version is the complement of that set. For the length split, we take the same approach using the length (in words) of the post, and the ID version of the subset is the set with posts of length less than the median length. For the relationships split, we take the ID version of the subset to be all posts in the r/relationships subreddit, and the OOD version to be the complement.

We apply this same splitting procedure to both the SFT and RM training datasets. Table 6 and Table 7 show the size of the training, validation, testing and OOD testing sets for the RLHF and SFT, and RM training, respectively. For the results in this work we randomly sample from the test and OOD test sets for those metrics, and we randomly sample from the validation set (which is in-distribution) for calculating metrics used for model selection.

Table 7: Size of the train, validation, test and OOD test datasets for each split for the reward models.

	length	sentiment	relationships
Train	45395	46411	52346
Validation	16513	16529	17687
Test	25539	25353	27492
OOD Test	25180	25366	23227

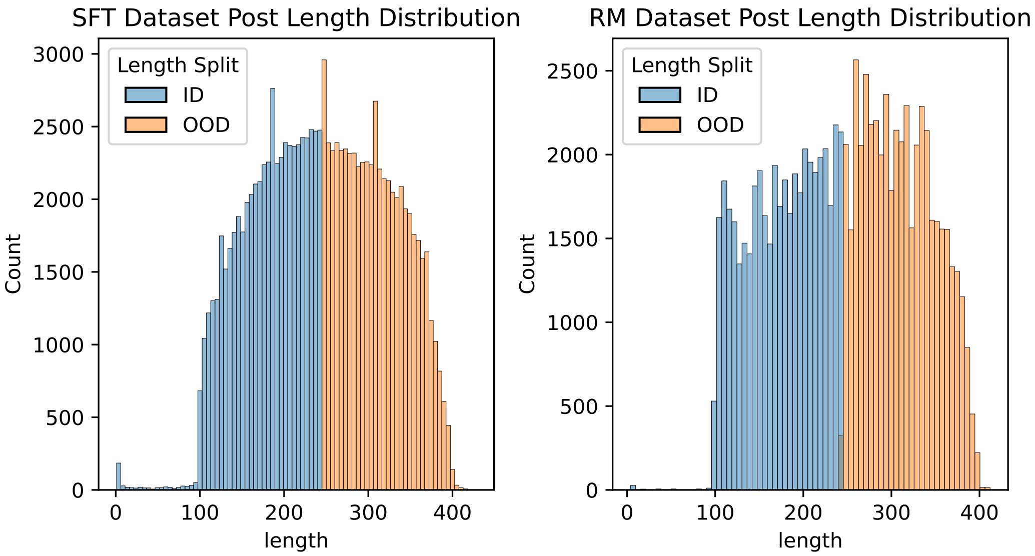

Figure 14: The distribution of post lengths across the full SFT and RM datasets. ID is the in-distribution version of the dataset, and OOD is the out-of-distribution version.

To understand the distribution shifts these splits entail, we show density plots for post length and sentiment across the full SFT and RM dataset in Fig. 14 and Fig. 15, and show the number of posts in each subreddit in Fig. 16.

J.2 Hyperparameters

We use the same hyperparameters as in LLaMa training (see Appendix E.4), but sweep over learning rates for each model size. We detail the learning rates swept over for each model size and the chosen learning rate, for SFT, RM and RLHF training in Table 8, Table 9 and Table 10 respectively. In general for SFT and RLHF we did not see much variance with seeds, but we did in RM training, matching prior work (Stiennon et al., 2022). For the RLHF training with the largest two model sizes, due to the large amount of compute required to run multiple training runs, for several model size and dataset shift combinations we chose a single learning rate based on what we thought would give the best results at the time we started training. For the combinations where we did vary the learning rate we did not see much variation in performance on the metrics we measured, so we do not expect these choices to affect the results.

Table 8: Learning rates for different model sizes and dataset splits for SFT 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	$\underline{1e^{-4}}$,$3e^{-5}$,$1.5e^{-5}$	$\underline{1e^{-4}}$,$3e^{-5}$,$1.5e^{-5}$	$\underline{1e^{-4}}$,$3e^{-5}$,$1.5e^{-5}$	$\underline{1e^{-4}}$,$3e^{-5}$,$1.5e^{-5}$	$1e^{-4}$,$3e^{-5}$,$\underline{1.5e^{-5}}$
length	$1e^{-4}$,$3e^{-5}$,$\underline{1.5e^{-5}}$	$\underline{1e^{-4}}$,$3e^{-5}$,$1.5e^{-5}$	$1e^{-4}$,$3e^{-5}$,$\underline{1.5e^{-5}}$	$1e^{-4}$,$3e^{-5}$,$\underline{1.5e^{-5}}$	$\underline{1e^{-4}}$,$3e^{-5}$,$1.5e^{-5}$
sentiment	$\underline{1e^{-4}}$,$3e^{-5}$,$1.5e^{-5}$	$1e^{-4}$,$\underline{3e^{-5}}$,$1.5e^{-5}$	$1e^{-4}$,$3e^{-5}$,$\underline{1.5e^{-5}}$	$\underline{1e^{-4}}$,$3e^{-5}$,$1.5e^{-5}$	$\underline{1e^{-4}}$,$3e^{-5}$,$1.5e^{-5}$