Reuse Your Rewards:
Reward Model Transfer for Zero-Shot Cross-Lingual Alignment

Zhaofeng Wu[MIT] **Ananth Balashankar**[DeepMind] **Yoon Kim**[MIT] **Jacob Eisenstein**[DeepMind] **Ahmad Beirami**[DeepMind]
[MIT]MIT   [DeepMind]Google DeepMind
zfw@csail.mit.edu

Abstract

Aligning language models (LMs) based on human-annotated preference data is a crucial step in obtaining practical and performant LM-based systems. However, multilingual human preference data are difficult to obtain at scale, making it challenging to extend this framework to diverse languages. In this work, we evaluate a simple approach for zero-shot cross-lingual alignment, where a reward model is trained on preference data in one source language and directly applied to other target languages. On summarization and open-ended dialog generation, we show that this method is consistently successful under comprehensive evaluation settings, including human evaluation: cross-lingually aligned models are preferred by humans over unaligned models on up to >70% of evaluation instances. We moreover find that a different-language reward model sometimes yields better aligned models than a same-language reward model. We also identify best practices when there is no language-specific data for even supervised finetuning, another component in alignment.

1 Introduction

Alignment has become an indispensable stage for building practical language models (LMs) adjusted to human preferences. This additional step, however, makes it challenging to develop LMs for many languages: unlike for autoregressive language modeling where multilingual unlabeled data may be easy to obtain (Joshi et al., 2020), such as religious texts (Christodouloupoulos and Steedman, 2015), labeled preference data can be expensive to gather. How do we align a LM in a target language without any preference data in that language?

We propose a novel reward model (RM) transfer setup, where we re-purpose a trained RM for some source language to align a LM in a target language (Figure 1), and investigate the effectiveness
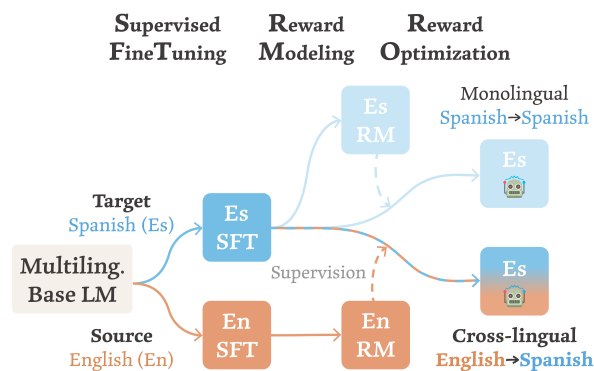
Figure 1: Cross-lingual reward model (RM) transfer. To align in a target language (in this example, Spanish), common monolingual alignment uses a RM for that target language. Instead, we re-purpose a RM for a different source language (in this example, English).
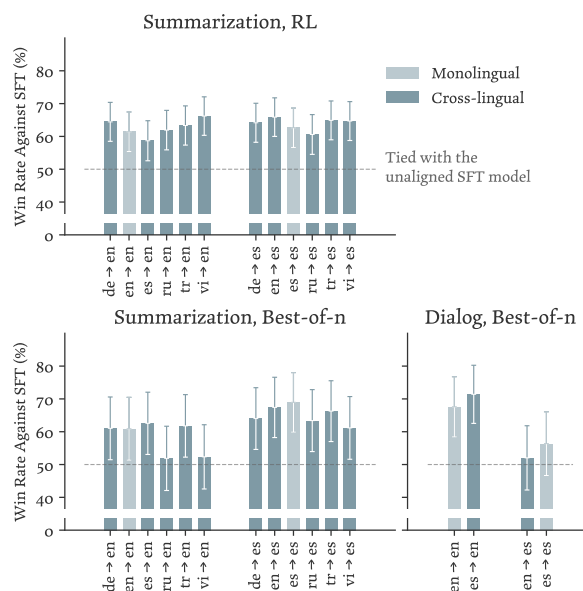
Figure 2: Performing target-language alignment using a RM for a different source language improves performance, when evaluated exclusively in the target language. This improvement is sometimes even larger than using the target-language RM (monolingual alignment). Here we measure the win rate against the target-language (unaligned) SFT model judged by humans, and the 95% confidence interval across validation instances. "source→target" denotes using a source-language RM to drive alignment in the target language.

Work done while ZW was a part-time intern at Google.

of this simple recipe. Across two tasks (summarization and open-ended dialog generation), two reward optimization methods (reinforcement learning and best-of-n reranking), and various evaluation settings, we demonstrate substantial and consistent zero-shot cross-lingual utility of RMs. Surprisingly, alignment using a different-language RM sometimes outperforms using a same-language RM, both when judged by humans and LMs. We also show that our RM transfer framework is useful even when target-language data for supervised finetuning (SFT), another component in alignment, is inaccessible.

Our results show that RM signals are generalizable and robust to input distribution changes, which could be leveraged for more future applications. Practically, our findings pave the path towards lowering the costs for training and deploying LMs that more equitably serve users around the world.

2 Background: Alignment From Human Feedback

In addition to traditional unsupervised LM pretraining, many recent LMs also include an alignment phase to improve helpfulness, harmlessness, etc., supervised by human feedback (Bai et al., 2022a; Ouyang et al., 2022; *i.a.*). A common recipe includes three stages: supervised finetuning (SFT), reward modeling (RM), and reward optimization. We give an overview of each and refer readers to Ouyang et al. (2022) and Bai et al. (2022a) for details. We assume a **base** model already pretrained using a usually next-token prediction objective.

The SFT stage initializes from the base model and takes task inputs $x \in \mathcal{X}$ to train the model to simulate example outputs $y \in \mathcal{Y}$. Specifically, it optimizes the conditional log-likelihood of y given some input x, similar to regular language modeling. We denote the trained SFT model using π_{SFT}.

The RM stage trains a model $r : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ as a proxy for human-judged quality of y under x. It initializes from π_{SFT} and is trained using a dataset of human judgments of generations. We consider two types of feedback to train the RM:

1. **Pointwise** feedback judges the quality of a single generation; in particular we only consider binary (good or bad) pointwise judgments. Denoting it as $z \in \{0, 1\}$ and letting $\mathcal{D}^{\text{RM}}$ be a dataset of judgments, the RM can be a standard classifier trained using the cross-entropy loss,

$$-\mathbb{E}_{(x,y,z)\sim\mathcal{D}^{\text{RM}}}[z \log \sigma\left(r(x, y)\right) + (1 - z) \log\left(1 - \sigma\left(1 - r(x, y)\right)\right)].$$

2. **Pairwise** feedback chooses a better generation out of two. We denote the chosen one as y_w and the other as y_l. To train a pointwise RM on such data, the Bradley-Terry model (Bradley and Terry, 1952) is often used, maximizing

$$\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}^{\text{RM}}}[\log \sigma\left(r(x, y_w) - r(x, y_l)\right)].$$

It is also generalizable to more than two outputs.

The reward optimization stage also initializes from π_{SFT} and further adjusts the model outputs using human feedback (as captured by the RM). Two common methods are reinforcement learning (RL) and best-of-n. Best-of-n is an inference-time procedure that does not change the underlying model, where multiple generations are sampled from π_{SFT} and then reranked using the RM; the highest-scoring generation is returned as the output. In RL, the model itself is changed such that its samples are scored highly by the RM, with the objective

$$\mathbb{E}_{x\sim\mathcal{D}^{\text{RO}},\tilde{y}\sim\pi_\theta(x)}[r(x, \tilde{y}) - \beta\left(\log \pi_\theta(\tilde{y} \mid x) - \log \pi_{\text{SFT}}(\tilde{y} \mid x)\right)].$$

$\mathcal{D}^{\text{RO}}$ is a dataset of inputs and β is a regularization hyperparameter. The above is typically optimized with PPO (Schulman et al., 2017). While we generally experiment with both methods, in some of our analyses we focus on best-of-n for a clean testbed without confounders from RL training.

3 Reward Model Transfer for Cross-Lingual Alignment

The pipeline in §2 is usually performed monolingually, commonly in English. Aligning for a new language requires both SFT data and RM data in that language. While the former may be relatively easier to obtain due to automatic construction methods, such as by re-purposing existing multilingual datasets (Muennighoff et al., 2023) or by eliciting from LMs (Wang et al., 2023c), RM data for a new language can be more expensive to gather, as it in principle requires human judgments. Additionally, RM data should ideally be periodically re-collected to avoid over-optimization (Bai et al., 2022a), further increasing data demand. Thus, we are mainly interested in alignment without target-language RM data, though, in §5.3, we investigate dispensing with target-language SFT data too.

We propose to perform reward optimization using a RM trained for a different language (Figure 1). Intuitively, assuming model generation quality transfers cross-lingually (e.g., good English generations are still good when translated into Spanish[1]), a model that can judge the output quality in one language should generalize to others, as long as the RM understands the languages, which is enabled by multilingual *base* model training. This generalizability is often observed for other tasks in the zero-shot cross-lingual transfer literature (Wu and Dredze, 2019; Pires et al., 2019; Conneau et al., 2020b; Hu et al., 2020; *i.a.*), and we expect it to work for RMs too. A simple baseline would be to use automatically translated RM data, to which we compare in §5.1. In this paper, we use **source language** to denote the RM language, and **target language** for the language of the aligned model.

4 Experimental Setup

We consider two tasks: summarization, common in alignment research (Stiennon et al., 2020; Ziegler et al., 2020; Lee et al., 2023; *i.a.*), and open-ended dialog generation, with substantial real-world relevance. §A describes dataset details and statistics. §B includes training details. §G.1 contains our task instructions.

Summarization. The Seahorse dataset (Clark et al., 2023) contains documents and summaries in six languages (German, English, Spanish, Russian, Turkish, and Vietnamese) with pointwise human ratings which we use. For SFT, we gather the data sources of Seahorse: XSum (Narayan et al., 2018), XL-Sum (Hasan et al., 2021), MLSum (Scialom et al., 2020), and WikiLingua (Ladhak et al., 2020). We use mT5-XL (Xue et al., 2021) as our multilingual base model, with 3.7B parameters.

Open-Ended Dialog Generation. We use the OpenAssistant dataset (Köpf et al., 2023) with multilingual, pairwise human-rated chat transcripts.[2] For the SFT data, we use the human-preferred response in each pair to finetune the model. Many languages in OpenAssistant have only limited data, so we only consider three languages with the most amounts of data: English, Spanish, and Russian.

We use PaLM-2-XXS as the base model (Anil et al., 2023). The authors of OpenAssistant found RL to be ineffective for this dataset (Köpf et al., 2023), which we confirmed in our experiments (Figure 4). We therefore focus on best-of-n for this task.

Evaluation. We assess model quality across several settings. First, we use the target-language RM, which is by design finetuned to judge target-language generation quality. But because of potential RM biases (Gao et al., 2023; Coste et al., 2023; Eisenstein et al., 2023), we also include two zero-shot-prompted evaluation models with much larger backbones—GPT-4 (OpenAI, 2023) and PaLM-2-L (Anil et al., 2023). This latter evaluation setup is common in prior work and has been demonstrated to correlate well with human judgments (Lee et al., 2023; Rafailov et al., 2023; An et al., 2023; Mu et al., 2023; *i.a.*). We also confirm its validity in §5.1 and §C. Importantly, both evaluation LMs support multilingual texts. Finally, we also perform human evaluations by self-reported native or advanced speakers, though only for a subset of language pairs and 250 (RL) / 100 (best-of-n) instances per pair due to its cost. For both human and LM evaluation, we elicit pairwise judgments to compare responses from the aligned model and the SFT model (Bai et al., 2022b; Lee et al., 2023; *i.a.*). We measure the *win rate*, i.e., how often the judge prefers the former. A 50% win rate indicates no improvement from alignment. §G.2 includes more details such as the evaluation prompts and positional bias control.

5 Results

Here we report the results of cross-lingual alignment. See §H for numerical results that correspond to the plots in this section.

5.1 Cross-Lingual Alignment Is Effective

When evaluated by the finetuned target-language RM, Figure 3 shows that monolingual best-of-n or RL always improves model quality, as expected. Encouragingly, cross-lingual reward optimization improves over the SFT model in all cases too. Similarly, when judged by a general-purpose LM, PaLM-2-L in Figure 4 and GPT-4 in §D, in-language and cross-lingual reward optimization both generally improve model quality. Importantly, we observe high agreement between the two LMs: on an instance level, they agree >70% across setups (see §D); if we consider how often they agree in

---

[1] We believe this is a weak assumption, though for tasks and instances more subject to culture-specific factors, generations may be judged more differently across languages (Costa et al., 2014; Hershcovich et al., 2022; Shwartz, 2022).

[2] In https://huggingface.co/datasets/OpenAssistant/oasst1.

the relative ranking of two source languages, they agree 78% for summarization (both best-of-n and RL) and 100% for dialog generation (best-of-n). This indicates the reliability of a LM judge.

Human evaluation (Figure 2) reveals the same trend, though with larger confidence intervals due to the cost. Human evaluation results also validate and justify LM-based evaluation: For summarization, PaLM-2-L (GPT-4) agrees with humans 65% (69%) of the time in English and 66% (62%) in Spanish, matching the 63% human-human agreement for English reference summaries and 67% for Spanish in Seahorse (Clark, personal communication, April 15, 2024). For dialog, PaLM-2-L (GPT-4) agrees with humans 69% (59%) of the time in English and 62% (60%) in Spanish, again similar to the 63% human-human agreement in Bai et al. (2022a) and 66% in Dubois et al. (2024). With further evidence in §C, we believe our LM judges reasonably reflect output quality.

We also compare our cross-lingual transfer setup to an alternative strategy, sometimes dubbed "translate-train" (Conneau et al., 2018; *i.a.*), that first trains a silver target-language RM by automatically translating the source-language data and then using the silver RM for target-language alignment. Averaged across all 30 ($= 6^2 - 6$) cross-lingual language pairs, under best-of-n and judged by PaLM-2-L, our RM transfer strategy outperforms translate-train[3] (average win rate 58.8 vs. 57.5; see Table 6 and 17 for raw numbers). RM transfer also has an efficiency advantage: to align in multiple target languages, it suffices to train one source-language RM, rather than different ones for each target language. In §F, we also explore alignment using bilingual RMs with two source languages (Mulcaire et al., 2019), though without noticeable improvements.

5.2 Cross-Lingual Alignment Sometimes Outperforms Monolingual Alignment

Remarkably, cross-lingual reward optimization often yields an even better model than using the target-language RM. This is validated by (1) the consistent trend when evaluated by PaLM-2-L, GPT-4, and humans, (2) their instance-level and ranking-level agreement (§5.1), and (3) the small confidence intervals. This may be due to a regularization effect: the target-language RM may possess language-specific spurious artifacts, to which the target-language policy model can overfit (Gao et al.,

---
[3]Which we implement using Google Translate.
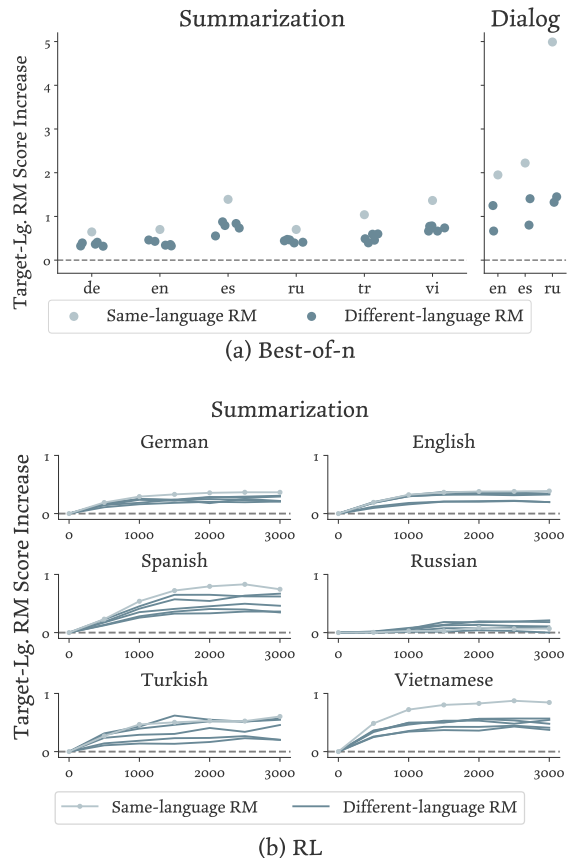

(a) Best-of-n

(b) RL

Figure 3: Cross-lingual alignment effectiveness judged by a finetuned target-language RM evaluator, measured in its score increase between the aligned model and the target-language SFT model. Each group in (a) and subplot in (b) represents one target language, and different dots/lines within each represent different source languages. RL is difficult to train for OpenAssistant (§4), so we omit it here. **In most cases, the RM evaluator score improves for cross-lingually aligned models.**

2023) more than artifacts in a different language in the source-language RM. Suppose, for example, that the target-language RM assigns higher rewards when the generation contains certain target-language words (due to bias in the RM training data). A different-language policy model is unlikely to exploit this, as it rarely generates these words, but a same-language policy model may.

This hypothesis is consistent with our observed patterns. First, there are many fewer cases of cross-lingual reward optimization outperforming the monolingual setting when measured by the finetuned target-language RM evaluator than the prompted LM evaluators (Figure 3): under this hypothesis, the finetuned evaluator RMs would be more susceptible to such artifacts and (incorrectly) assign higher scores in the monolingual settings. The underperformance of the translate-train baseline (§5.1) also provides weak evidence: in princi-
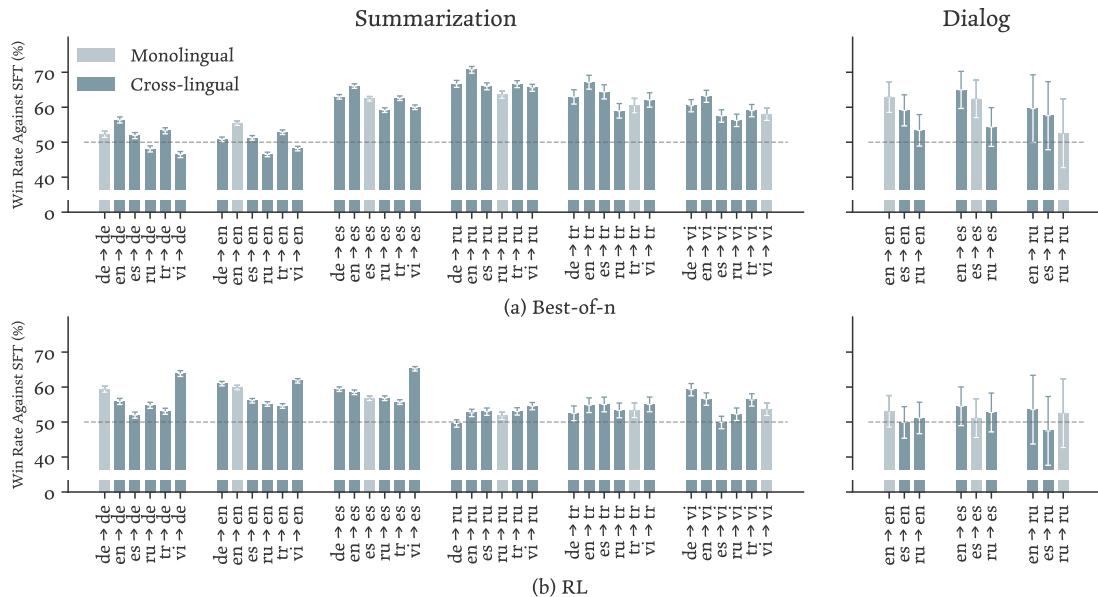
Figure 4: Alignment effectiveness, compared to the target-language SFT model judged by PaLM-2-L, and the 95% confidence interval across validation instances. "source→target" denotes a source-language RM driving alignment in the target language. **Cross-lingual alignment is generally effective, sometimes outperforming monolingual alignment**. RL is hard to train for OpenAssistant, in line with what its authors found (Köpf et al., 2023).

ple, a source-language RM and a source-translated-into-target-language RM should capture the same reward signal, as they are derived from the same data source, and would lead to similar downstream performance. However, the former is less susceptible to reward over-optimization due to the language mismatch, leading to better performance, though this is confounded by translation quality.

Corroborating this hypothesis, we also find that when used monolingually, the RMs behave more like a bag-of-word (BoW) model. We take each of the 6 summarization RMs and infer on the validation set of each dataset in each language (Table 1). In every setting, we fit a BoW linear regressor to predict the RM-assigned score for each instance and compute the R^2 across instances as a proxy for the RM's similarity to a BoW model in that setting. For each dataset, and for every source language that differs from the dataset's language, we check whether inferring using the source-language RM or the dataset-language RM results in a larger R^2. The latter monolingual usage has a higher R^2 (0.65 vs. 0.63), so it is more likely that the RMs overfit to lexical patterns when used in-language.

5.3 Cross-Lingual Alignment Without Target-Language SFT Data

So far we assumed access to target-language SFT data since, as §3 argues, SFT data could be more easily obtained than RM data. We now relax this assumption and instead translate the source-language

SFT data into the target language using Google Translate. We investigate if it, combined with RM transfer, still enables cross-lingual alignment. As a case study, we only consider summarization and when English is the source or target language.

Using translated SFT data substantially degrades the quality of the SFT model (Figure 5(a)) and the best-of-n-aligned LM (Figure 5(b)). There are however two factors: (1) quality loss due to translation, and (2) domain/style mismatch. For (2), we note that different languages have SFT data composed of different datasets, following Seahorse (Table 1).[4] And these datasets differ stylistically: for example, while XSum includes news articles, WikiLingua consists of how-to articles and with more formulaic summaries. There would thus be a domain difference between using organic target-language SFT data vs. data translated from a different language.

To account for this, we employ round-trip back-translation, first translating the target-language SFT data into the source language and then back to the target language. This setup is not practically useful but it upper-bounds the effect of translation errors alone. Figure 5(a) shows that this bridges most of the gap, sometimes leading to models that win over the SFT model >50% of the time. Alternatively, we control for domain by repeating our experiments solely using WikiLingua for both SFT and RM as

---

[4]SFT data quantity may also be a confounder, but we consider directions both from and to English, and the degradation is substantial in both. So quantity is not the biggest factor.
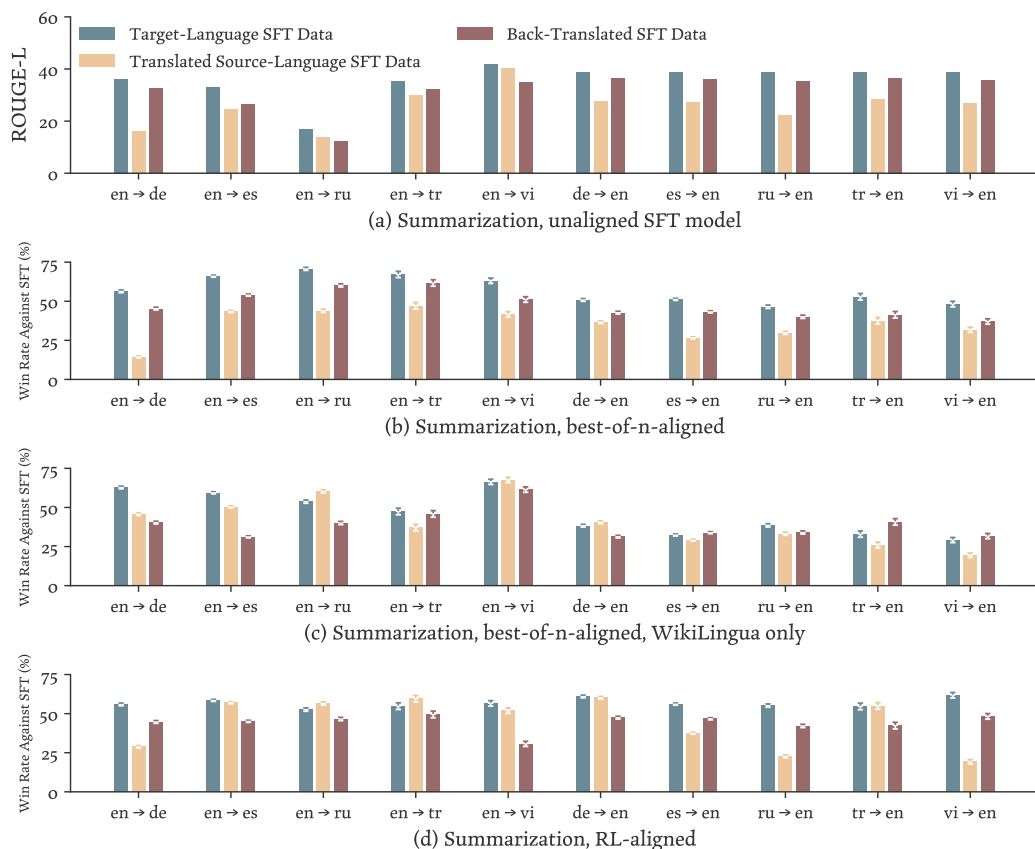
Figure 5: Cross-lingual alignment results without target-language SFT data using various strategies and on different data. **Training the SFT model using data translated from another language can be helpful when aligning using RL ((d)), but domain match is important for best-of-n** ((c) and the back-translation results).

it is present for all languages. From Figure 5(c), the gap indeed reduces, with the translated SFT models sometimes even outperforming the original, and back-translation is no longer consistently beneficial.

Other than genre control, we also hypothesize that the gap would be smaller for RL than best-of-n because the RM, whose transferability we verified (§5), intuitively plays a bigger role in the RL pipeline. Best-of-n, on the other hand, is more reliant on the SFT model quality, as reflected by the high resemblance between the transfer performance patterns in Figure 5(b) and the SFT model quality in Figure 5(a). Figure 5(d) indeed shows that the translated models have little performance drop, except for cases where the former degenerates.[5] Again, apart from the degenerate cases, back-translation is not helpful.

To summarize,[6] cross-lingual alignment could still be helpful even without target-language SFT data, though care needs to be taken when training

the surrogate SFT model. While we only experimented on summarization, we believe there will be larger text diversity for dialog generation in the wild, for which this issue warrants greater attention.

5.4 Practical Recommendations

Our findings suggest that, for SFT, it is always beneficial to use organic target-language data, but when inaccessible, automatic translation may be a remedy, though one should be mindful of the data distribution match between the data source and the application, or relying more on RL.

For RM, cross-lingual transfer is often successful, but how does one select the source RM language to align in a new target language? In Figure 6, we show the source languages ranked by transfer effectiveness for each target language. The rankings across target languages are generally stable, especially for best-of-n: if a source language is effective for one target language, it is usually effective for others too. Therefore, one may select the source language by extrapolating from its performance on other target languages. In particular, English RMs are usually the most accessible in
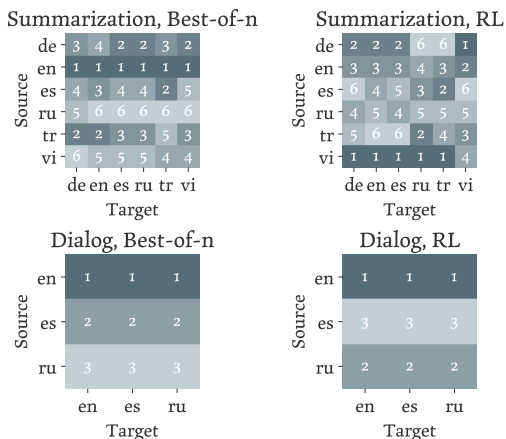
---

[5]Which we believe is due to a lack of careful case-by-case hyperparameter tuning, which we did not perform as it would be very expensive to tune for each transfer pair.

[6]No pun intended.

Figure 6: PaLM-2-L-judged rankings of source language effectiveness when driving alignment in different target languages. English is generally a good source.

practice. Our results show that it is a decent strategy to use them as the source: English is often a highly-ranked source language, most frequently the best, perhaps due to the relatively higher annotator quantity and quality (Yu et al., 2022) or implicit modeling assumptions (Dyer et al., 2019). Beyond this empirical observation, we try to causally predict the pairwise transferability from various features in §6, but without success.

6 Analysis

The effectiveness of cross-lingual alignment motivates us to better understand how it relates to various factors. We show that while RM generalizability within the original reward modeling task is a prerequisite, it does not uniquely explain the downstream success. Similarly, we also show that the pairwise win rates (judged by PaLM-2-L unless otherwise mentioned) cannot be fully explained by, and thereby not predictable from, language features or the KL-divergence from the SFT model.

6.1 Impact of RM Generalizability Within Reward Modeling

The RMs' cross-lingual utility in downstream alignment is predicated on their generalizability within the original reward modeling task, but the latter is not sufficient for the former. So how much does this generalizability explain the alignment success? We analyze this generalizability following the cross-lingual transfer tradition, zero-shot applying a source-language RM to the target-language validation data and computing accuracy (Wu and Dredze, 2019, 2020; Pires et al., 2019; *i.a.*). We also consider a majority baseline and a length baseline to check if the RMs are only superficially cap-

turing generation length (Wang et al., 2023b; Singhal et al., 2023). To compute this length baseline: for dialog generation, a pairwise task, all longer, or shorter, responses in each pair are chosen, depending on which (long or short) yields higher training set accuracy. For summarization, a pointwise task, all responses longer (or shorter) than a threshold are chosen. The direction (long or short) and the threshold are also selected using the training set.

Figure 7 confirms cross-lingual RM generalizability: cross-lingual RMs often perform above the majority baseline for summarization and random performance (50%) for dialog. §E verifies this cross-lingual generalizability with another setup.

Nevertheless, the improvements over the majority/random baselines are modest. The dialog models even sometimes underperform the length baseline (though this does not mean the RMs only rely on length[7]). Part of this is due to the high subjectivity of the reward modeling task: the RM accuracies here are near the human agreement level for Seahorse (Clark et al., 2023), plotted in Figure 7, and generally match the human agreement numbers in dialog generation work (Bai et al., 2022a; Dubois et al., 2024). But it is still interesting that seemingly weak RMs, like the Vietnamese RM which performs similarly to the majority baseline when used monolingually or the dialog RMs which are often surpassed by the length baseline, can achieve high cross-lingual alignment effectiveness (Figure 4).

Furthermore, the results here do not match their downstream utility, regardless of whether we consider the quality of the RMs as measured by their in-language validation accuracy (Turkish, for example, is the best in Figure 7, but not so in Figure 6), the generalizability of the RMs which we operationalize as the difference between in-language training and validation loss (or accuracy—they yield the same ranking: Russian, German, English, Turkish, Vietnamese, and Spanish, from the least amount of overfitting to the most, again different from Figure 6), or the specific pairwise transfer effectiveness (for each target language, we compare the effectiveness of source languages ranked by the reward modeling task generalizability here vs. by downstream alignment win rate; on summarization, averaged across target languages, Kendall's $\tau = 0.1$ (same with best-of-n or RL), indicat-

---

[7]The RMs agree with the length baseline on 72.6% of the validation instances, higher than the baseline agreement level of 56.6% (how often two random models at their accuracy levels agree on average), but far from full agreement.
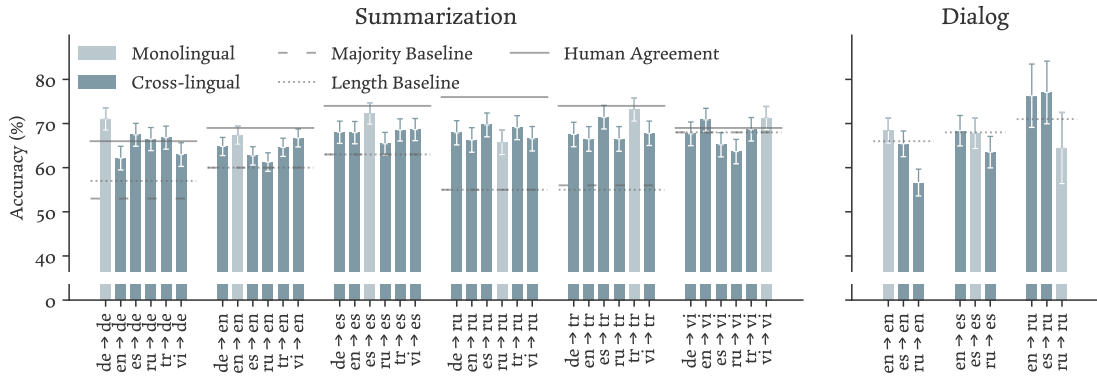
Figure 7: Source-language RM generalizability within the original reward modeling task and the 95% confidence interval across validation instances. "source→target" denotes training a source-language RM and measuring its accuracy on the target language validation data. The baselines are explained in §6.1. Dialog generation, a pairwise task, does not have a majority baseline; the dataset authors also did not report human agreement. **RMs generally exhibit cross-lingual generalizability, exceeding the majority baseline and often the length baseline**.

ing low ranking agreement). Overall, while cross-lingual alignment depends on RM generalizability on the original task, other factors are at play too.

6.2 Impact of Language Features

Can the cross-lingual alignment performance be predicted from simple language features, such as their frequency in the pretraining corpus or typological similarity? The summarization languages ranked by frequency in the mT5 corpus, the base model for this task, are: English, Russian, Spanish, German, Turkish, Vietnamese (Xue et al., 2021). This does not match the transfer utility ranking in Figure 6. Similarly, neither does the ranking match the SFT data quantity or RM data quantity (in §A).

Linguistic typology and orthography are also common predictors of cross-lingual transferability (Gerz et al., 2018; K et al., 2020; Muller et al., 2021; *i.a.*). This, however, is not the case for us either: for summarization RL, for example, English benefits from Vietnamese the most, but they belong to disparate language families. Orthography may be playing a role: Russian overall does not transfer well to other languages, and it is the only language that does not use the Latin script, but this trend is not clear. Systematically, we compute the correlation between alignment utility and WALS features of linguistic typology (Dryer and Haspelmath, 2013). For each WALS feature present for all 6 summarization languages, we divide all win rates into two groups: those between language pairs that have the same, or different, feature values. Under a one-sided unpaired t-test, no feature shows statistical significance at $\alpha = 0.05$ with Bonferroni correction (Dunn, 1961).[8] Therefore, alignment

---

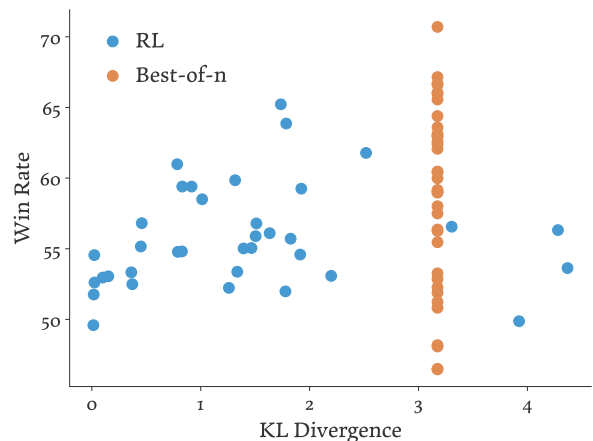[8] Even without correction, only 4 show statistical signifi-

Figure 8: Win rate (PaLM-2-L-judged) vs. KL-divergence for summarization across different (source, target) language pairs. For best-of-n, we use the upper bound formula in Stiennon et al. (2020), Beirami et al. (2024), *i.a.*, which is a function of n and thus appears as a vertical line. **KL-divergence does not fully explain the final alignment performance**.

utility does not strongly correlate with such language features.

6.3 Impact of Policy Divergence

From a learning angle, it has been shown that the reward that a learned policy can obtain strongly correlates with its KL-divergence from the base (SFT) policy (Bai et al., 2022a). This could be concerning, if the model deviates from the base policy to "hack" the reward (Gao et al., 2023; Coste et al., 2023; Eisenstein et al., 2023), but not if the evaluation metric is robust. As we perform human evaluation and also verified that our LM judges correlate with human judgments, this is less of a

cance at $\alpha = 0.05$ out of 123: 1A, 3A, 37A, and 54A. The first two are phonological features, and the other two minor syntactic features, thus likely being spurious correlations.

problem for us. Nevertheless, in Figure 8, we plot the correlation between the win rates and the KL-divergence of the aligned models. There is not a clear correlation, and hence we do not observe reward over-optimization.

7 Related Work

Zero-shot cross-lingual transfer. There is a long line of research on cross-lingual representation generalizability, such as with sentence embeddings (Conneau et al., 2018) or more recently, LMs (Wu and Dredze, 2019, 2020; Pires et al., 2019; Siddhant et al., 2020). Commonly, a multilingual LM (Devlin et al., 2019; Conneau and Lample, 2019; Conneau et al., 2020a; *i.a.*) is finetuned on a task in a source language and evaluated on the task's test set in a different language. This is generally effective. Our RM transfer setup can be viewed under this framework, but we go further and show that this generalizability is useful for *downstream* tasks, in our case alignment. Shaham et al. (2024) and Chirkova and Nikoulina (2024) are close to us in studying cross-lingual generalizability in alignment, but only focusing on SFT and only using translated data.

Multilingual Alignment. For SFT, it is common to assemble existing multilingual task datasets into instruction datasets (Muennighoff et al., 2023; Asai et al., 2023; Ahuja et al., 2023). Some have directly collected SFT data for non-English languages, either on a per-language basis (Zhang et al., 2023; Xu et al., 2023b; Ni et al., 2023; *i.a.*) or multilingually (Zhao et al., 2024; Singh et al., 2024), though this can be expensive. Past work has also used automatic translation for SFT (Li et al., 2023a; Lai et al., 2023; Shaham et al., 2024; *i.a.*) and RM data (Lai et al., 2023; Shen et al., 2024). We also use translation for SFT, but showed that cross-lingual transfer outperforms translation for RM.

8 Conclusion

We showed through two different tasks that we can perform alignment using a different-language RM. Surprisingly, we find this to be sometimes more effective than using a same-language RM. We also identified issues and remedies when we dispense with target-language SFT data. We hope our findings can motivate future work to build better LMs for more languages. Adapting our RM transfer setup to other settings such as domain generalization would also be exciting future directions.

Limitations

Free-form generation is challenging to evaluate, especially in a cross-lingual setup. As we mentioned, neither the finetuned target-language RM evaluator scores nor pairwise evaluation from humans or LMs are perfect (Wang et al., 2023b; Zheng et al., 2023; Hosking et al., 2024; *i.a.*). Nevertheless, we believe the consistent cross-lingual transferability observed across our many evaluation settings suggests that it would hold more generally. Similarly, it is not possible to comprehensively study the myriad of reward optimization methods (Rafailov et al., 2023; Azar et al., 2023; *i.a.*), some of which may not enjoy the same cross-lingual RM transfer benefit (in fact, the notion of a RM do not even exist in some, though analogous ideas may be applicable). However, the two that we study, best-of-n and PPO, are representative of current common practices, especially given the strong empirical performance of best-of-n (Gao et al., 2023; Mudgal et al., 2023; Rafailov et al., 2023; *i.a.*). Somewhat orthogonally, past work has argued that it is limiting to use one single scalar to represent generation quality (Xu et al., 2023a; Krishna et al., 2023; Hosking et al., 2024) and that more fine-grained rewards could be beneficial (Wu et al., 2023). We follow the convention to use one single score to more easily measure and compare cross-lingual transfer in many setups, but a similar but more fine-grained study would be valuable future work. It has also been shown that it is more challenging to train reward models for low-resourced languages (Shen et al., 2024). We only considered relatively high-resourced languages in this work, and it is possible that the pattern would differ when using lower-resourced source languages for transfer. Finally, our motivating assumption that generation quality being language-agnostic does not always hold, especially when facing culture-specific tasks or task instances. In those cases, we believe we would see reduced cross-lingual generalizability.

Acknowledgments

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Chenxin An, Shansan Gong, Ming Zhong, Xingjian Zhao, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-Eval: Instituting standardized evaluation for long context language models.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 technical report.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu,

Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional AI: Harmlessness from AI feedback.

Ahmad Beirami, Alekh Agarwal, Jonathan Berant, Alexander D'Amour, Jacob Eisenstein, Chirag Nagpal, and Ananda Theertha Suresh. 2024. Theoretical guarantees on the best-of-n alignment policy.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language model.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Elizabeth Clark, Shruti Rijhwani, Sebastian Gehrmann, Joshua Maynez, Roee Aharoni, Vitaly Nikolaev, Thibault Sellam, Aditya Siddhant, Dipanjan Das, and

Ankur Parikh. 2023. SEAHORSE: A multilingual, multifaceted dataset for summarization evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9397–9413, Singapore. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Albert Costa, Alice Foucart, Sayuri Hayakawa, Melina Aparici, Jose Apesteguia, Joy Heafner, and Boaz Keysar. 2014. Your morals depend on language. *PLOS ONE*, 9(4):1–7.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. 2023. Reward model ensembles help mitigate overoptimization.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2024. Alpaca-farm: A simulation framework for methods that learn from human feedback.

Olive Jean Dunn. 1961. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64.

Chris Dyer, Gábor Melis, and Phil Blunsom. 2019. A critical analysis of biased parsers in unsupervised parsing.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, Peter Shaw, and Jonathan Berant. 2023. Helping or herding? Reward model ensembles mitigate but do not eliminate reward hacking.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327, Brussels, Belgium. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human feedback is not gold standard.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations*.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. OpenAssistant conversations – democratizing large language model alignment.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. RLAIF: Scaling reinforcement learning from human feedback with ai feedback.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-X: Multilingual replicable instruction-following models with low-rank adaptation.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jesse Mu, Xiang Lisa Li, and Noah Goodman. 2023. Learning to compress prompts with gist tokens.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, YaGuang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, Jilin Chen, Alex Beutel, and Ahmad Beirami. 2023. Controlled decoding from language models.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3912–3918, Minneapolis, Minnesota. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 448–462, Online. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Jinjie Ni, Fuzhao Xue, Yuntian Deng, Jason Phang, Kabir Jain, Mahir Hitesh Shah, Zangwei Zheng, and Yang You. 2023. Instruction in the wild: A user-based instruction dataset. https://github.com/XueFuzhao/InstructionWild.

OpenAI. 2023. GPT-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. The language barrier: Dissecting safety challenges of llms in multilingual contexts.

Vered Shwartz. 2022. Good night at 4 pm?! Time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Ari, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2020. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8854–8861.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in RLHF.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023b. How far can camels go? Exploring the state of instruction tuning on open resources. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023c. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023a. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Guohai Xu, Jiayi Liu, Ming Yan, Haotian Xu, Jinghui Si, Zhuoran Zhou, Peng Yi, Xing Gao, Jitao Sang, Rong Zhang, Ji Zhang, Chao Peng, Fei Huang, and Jingren Zhou. 2023b. CValues: Measuring the values of chinese large language models from safety to responsibility.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. 2022. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3725–3743, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ge Zhang, Yemin Shi, Ruibo Liu, Ruibin Yuan, Yizhi Li, Siwei Dong, Yu Shu, Zhaoqun Li, Zekun Wang,

Chenghua Lin, Wenhao Huang, and Jie Fu. 2023. Chinese open instruction generalist: A preliminary release.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. (InThe)WildChat: 570k ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences.

A Dataset Details and Statistics

We report dataset statistics in Table 1, 2, 3, and 4. We reuse the SFT data for reward optimization (for both training and evaluation for RL, and for only evaluation for best-of-n since it does not have a training stage), but only the input x, without reference generations y.

The summarization SFT datasets, reported in Table 1, are the original data sources of Seahorse, which we take from the GEM release (Gehrmann et al., 2021). They are evenly mixed at the instance level for both SFT training and RL training. For evaluation of the aligned model, we macro-average the per-dataset metrics (e.g., win rate) for a language-level score. Because the Seahorse dataset was created using the validation and test instances of the original summarization datasets, to be clean, we exclude the Seahorse training instances from these splits when performing SFT and reward optimization. OpenAssistant does not have this issue and has clean split separations. The Seahorse summaries are human-rated along six axes, and we only use the sixth axis for our pointwise reward as it encapsulates previous axes (Clark et al., 2023). We limit the maximum length of model inputs to 1,024 tokens and outputs to 512 tokens. See also §G.1 for instructions we attach to the dataset instances during training and inference.

B Training Details

SFT. The model is trained using Adafactor (Shazeer and Stern, 2018) with a constant learning rate at 10^{-3} for summarization and 10^{-5} for dialog generation, batch size 32, and dropout 0.1. We perform checkpoint selection using validation ROUGE-L score (Lin, 2004).

RM. The model is trained using Adafactor with a constant learning rate at 10^{-4} after 1,000 linear warm-up steps, batch size 32, and dropout 0.1. We perform checkpoint selection using validation loss.

RL. We use PPO for RL training with a constant learning rate at 10^{-4}, batch size 32, for 3,000 steps for summarization and 2,500 steps for dialog generation. The value model has 1,000 linear warm-up steps and we only start training the policy model after 2,000 steps elapse. We set the regularization coefficient at $\beta = 0.1$.

Best-of-n. We use $n = 64$.

		Train	Validation
German	MLSum	220748	8932
	WikiLingua	40839	3699
English	XSum	23206	642
	XL-Sum	306522	9690
	WikiLingua	99020	12021
Spanish	XL-Sum	38110	3170
	MLSum	259888	8374
	WikiLingua	79212	9730
Russian	XL-Sum	62243	5492
	WikiLingua	37028	3209
Turkish	XL-Sum	27176	1953
	WikiLingua	3148	194
Vietnamese	XL-Sum	32111	2341
	WikiLingua	13707	679

Table 1: Number of summarization instances for the SFT and reward optimization stages. The datasets are taken from the GEM release (Gehrmann et al., 2021) and with certain validation instances removed (§A).

	Train	Validation
German	8389	1250
English	14031	2071
Spanish	8741	1310
Russian	7679	1112
Turkish	7855	1096
Vietnamese	7844	1166

Table 2: Number of summarization instances for reward modeling.

	Train	Validation
English	8898	472
Spanish	5681	311
Russian	1884	99

Table 3: Number of dialog generation instances for the SFT and reward optimization stages.

	Train	Validation
English	22076	1026
Spanish	13714	699
Russian	2627	135

Table 4: Number of dialog generation instances for reward modeling.

		De	En	Es	Ru	Tr	Vi
Summarization	Acc.	73.5%	73.0%	73.2%	73.7%	73.6%	78.2%
	N	306	1672	295	255	720	349
Dialog	Acc.	–	72.0%	70.8%	73.3%	–	–
	N	–	472	311	99	–	–

Table 5: The accuracy of evaluating the PaLM-2-L judge on the RM validation data. We also report the number of comparisons based on which the accuracy is calculated.
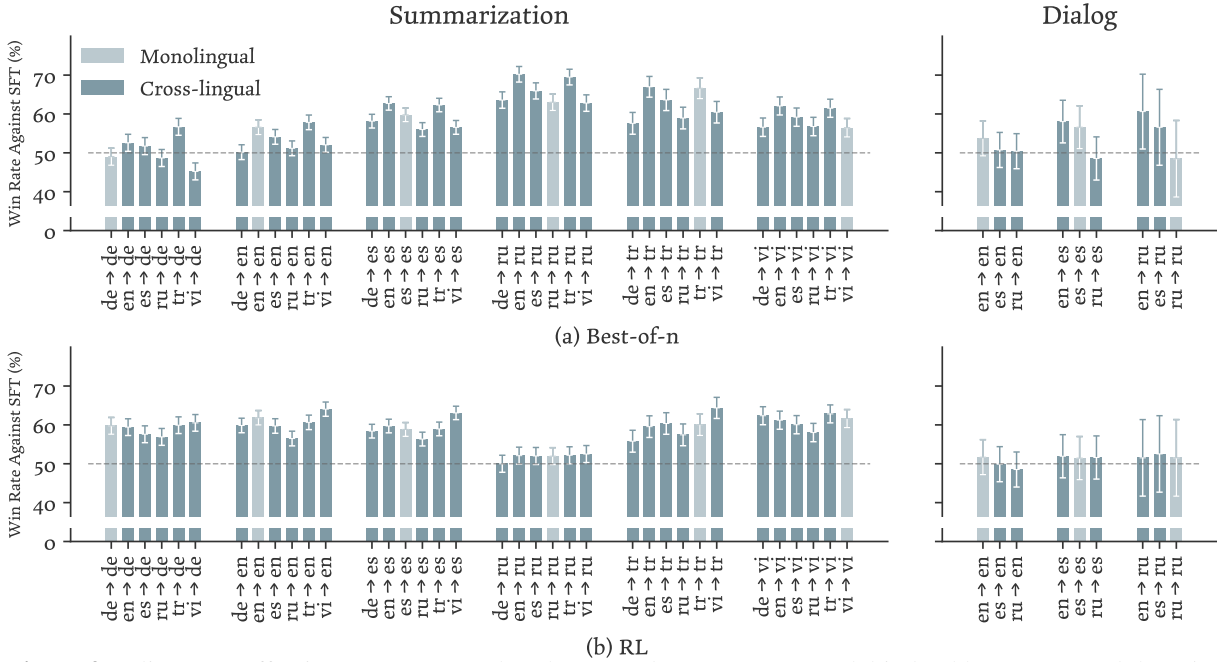

(a) Best-of-n

(b) RL

Figure 9: Alignment effectiveness, compared to the target-language SFT model judged by GPT-4, and the 95% confidence interval across validation instances. "source→target" denotes a source-language RM driving alignment in the target language. **Cross-lingual alignment is generally effective, sometimes outperforming monolingual alignment**. RL is hard to train for OpenAssistant, in line with what its authors found (Köpf et al., 2023).

C LM Judge Accuracy on Ground-truth Reward Modeling Data

We verify the validity of using LM as a judge for our tasks by computing its accuracy on the validation splits of the RM datasets we used. We only consider PaLM-2-L as a case study. For OpenAssistant, a pairwise dataset, we simply check if the RM ranks the candidate generations correctly according to human preference. For Seahorse, a pointwise dataset, we group summaries for the same source document, and for each summary pair in such groups, we compute the ranking correctness.

We show the results in Table 5. The accuracies generally match the human agreement in Seahorse (Clark et al., 2023), and while human agreement was not reported in OpenAssistant, they generally match the human agreement numbers in past work on dialog generation (Bai et al., 2022a; Dubois et al., 2024) too (see §5.1 for reference hu-

man agreement numbers). Taken together with the LM judges' agreement with human evaluation (§5.1), we believe it is valid to use a LM to assess the generation quality in our setup.

D GPT-4 as a Judge Results

In this section, we present the alignment evaluation results as judged by GPT-4, specifically the `gpt-4-0125-preview` model. Due to its high cost, we cap the number of evaluation instances for each dataset at 1,000 (i.e., for each row of Table 1 and 3). The results are shown in Figure 9. We observe the same trends as in §5.1, where cross-lingual reward optimization is generally effective, sometimes even more so than when done monolingually. Compared to PaLM-2-L, the two LMs agree on 72% of the instances in English and 71% in Spanish for summarization, and 75% and 73% for these languages for dialog. These are higher than the
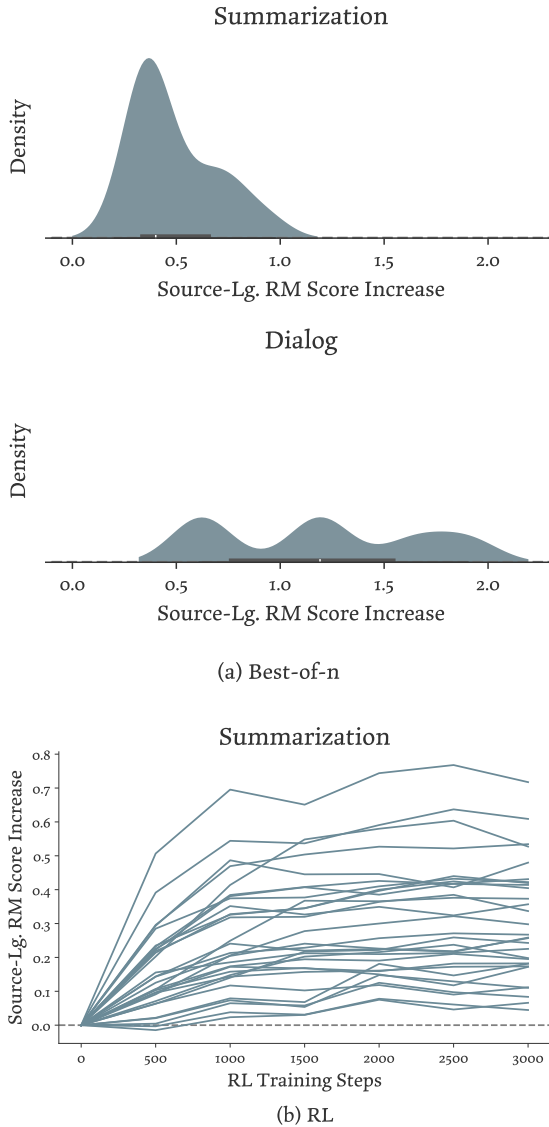
Figure 10: Source-language RM generalizability evaluated by increases in scores they assign to target-language generations after monolingual target-language alignment (best-of-n or RL). We show all (source, target) language pairs where the two languages differ as density in (a) and lines in (b). RL is difficult to train for OpenAssistant (§4), so we omit it here, since the assumption that the RL'ed model is better would not hold. **In most cases, the source-language RM assigns a higher score (>0 increase) to aligned models, demonstrating cross-lingual RM generalizability.**

baseline human-human agreement numbers in §5.1. This shows a sign of homogeneity between LM judges, but also confirms their reliability.

E Verifying RM Transfer for Reward Modeling

In §6.1, we observed RM generalizability on the original reward modeling task, which would be a necessary condition for successful downstream

cross-lingual alignment. There, we showed that the source-language RMs assign higher scores to better target-language generations than worse ones. Here, we consider an alternative setup to study the same problem: instead of relying on existing RM datasets for the better and worse generations, we take generations from monolingually-aligned models as better ones than those from unaligned (i.e., SFT) models. The assumption here is that monolingual alignment improves model quality, which is indeed the case as illustrated in Figure 4 and 9. Like in §6.1, we indeed see from Figure 10 that source-language RMs assign higher scores to monolingually-aligned models than unaligned SFT models. Under RL, this score difference also increases throughout training. These results confirm the RMs' cross-lingual generalizability within the reward modeling task.

F Alignment Using Bilingual RMs

Seeing the benefit of cross-lingual RM transferability in §5, we hypothesize that bilingual RMs could bring further improvements since the resulting reward could be encouraged to be more language agnostic (Mulcaire et al., 2019). It would be computationally expensive to experiment with all possible language configurations (there would be a cubic number of them with pairwise sources), so, for simplicity, we take the best-performing source languages under the summarization best-of-n setup as judged by PaLM-2-L, English and German (Figure 6), and see if a bilingual RM based on them would lead to further performance improvement. Specifically, we first train a bilingual SFT model by pooling the SFT data for both languages, and similarly for the RM, which initializes from this bilingual SFT model.

Figure 11 does not show an improvement from the bilingual RM, which always achieves similar performance to the English RM, the better of the two monolingual RMs. Nevertheless, if this trend holds consistently, that the bilingual RM matches the performance of the better monolingual RM, this could be useful as an alternative to having to perform source language selection. We leave a more systematic validation of this phenomenon to future work.

G Prompts

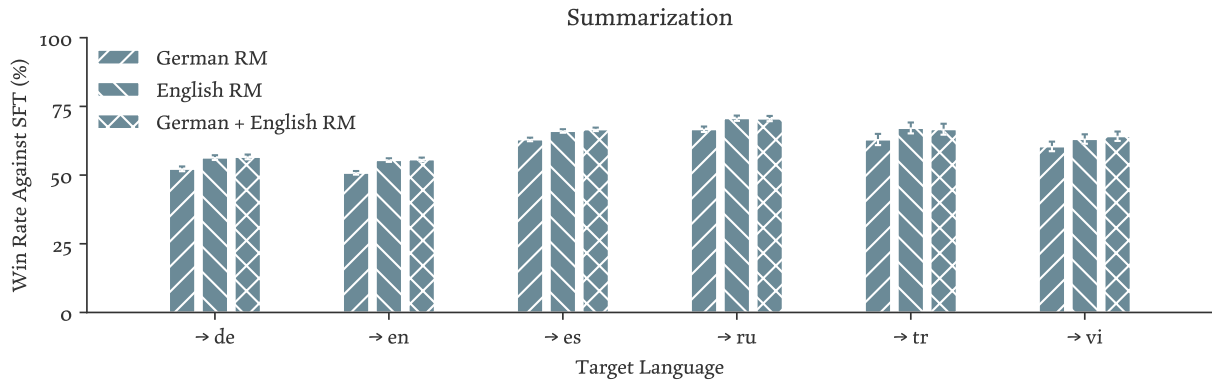In this section, we list all the prompts we used.

Figure 11: Alignment performance, measured in the win rate against the monolingual target-language SFT model, when alignment is driven by a German RM, an English RM, or a bilingual German + English RM. **The bilingual RM does not yield a noticeable improvement.**

G.1 Task Instructions

We prepend the following task-specific instructions to inputs for SFT and reward optimization. All occurrences of [LANGUAGE] are substituted with the target language. The RM stage does not include such prompts, where we simply concatenate the texts with delimiters.

Summarization: `Summarize the following text in [LANGUAGE]:`

Dialog generation: `You are given a dialog between a human and an assistant in [LANGUAGE]. Please write one turn of the assistant side in [LANGUAGE].\n\n`"

G.2 Evaluation Prompts

We use the following prompts to elicit pairwise generation judgments for both human and LM judge evaluation. All occurrences of [LANGUAGE], [INPUT], [GENERATION1], and [GENERATION2] are substituted with the respective content. For both tasks, we compare the probability of the tokens "1" and "2". To control for the positional bias of LMs (Wang et al., 2023a; Pezeshkpour and Hruschka, 2023; Zheng et al., 2023) and potentially of our human annotators, we randomly shuffle the two generations for human evaluation and the GPT-4 judge. For the PaLM-2 judge for which we have probability access, we prompt the LM judge twice with both orderings of the generations and compute the accuracy by averaging the probabilities of the "1" and "2" tokens.

Summarization. This prompt is adapted from the one in Lee et al. (2023).

```
A good summary is a shorter piece of text
that has the essence of the original. It
tries to accomplish the same purpose and
```

```
conveys the key information from the
original post. Below we define four
evaluation axes for summary quality:
coherence, accuracy, coverage, and overall
quality.

Coherence: This axis answers the question
"how coherent is the summary on its own?"
A summary is coherent if it's easy to
understand when read on its own and free of
English errors. A summary is not coherent
if it's difficult to understand what the
summary is trying to say. Generally, it's
more important that the summary is
understandable than it being free of
grammar errors.

Accuracy: This axis answers the question
"does the factual information in the
summary accurately match the post?" A
summary is accurate if it doesn't say
things that aren't in the article, it
doesn't mix up people, and generally is
not misleading.

Coverage: This axis answers the question
"how well does the summary cover the
important information in the post?" A
summary has good coverage if it mentions
the main information from the post that's
important to understand the situation
described in the post. A summary has poor
coverage if someone reading only the
summary would be missing several important
pieces of information about the situation
in the post. A summary with good coverage
should also match the purpose of the
```

original post (e.g. to ask for advice).

Overall quality: This axis answers the
question "how good is the summary overall
at representing the post?" This can
encompass all of the above axes of quality,
as well as others you feel are important.
If it's hard to find ways to make the
summary better, the overall quality is
good. If there are lots of different ways
the summary can be made better, the overall
quality is bad.

You are an expert summary rater and are
knowledgeable in [LANGUAGE]. Given a
piece of text in [LANGUAGE] and two of its
possible summaries, also in [LANGUAGE],
output 1 or 2 to indicate which summary
best adheres to coherence, accuracy,
coverage, and overall quality as defined
above.

Text - [INPUT]
Summary 1 - [GENERATION1]
Summary 2 - [GENERATION2]

Preferred Summary=

Dialog Generation This prompt is adapted from
the one in Li et al. (2023b).

You are a helpful assistant, that ranks
models by the quality of their answers.
You are also knowledgeable in [LANGUAGE].

I want you to create a leaderboard of
different large-language models. To do
so, I will give you the instructions
(prompts) given to the models, and the
responses of two models. Please rank the
models based on which response would be
preferred by humans. All inputs are
python dictionaries.

Here is the prompt, in [LANGUAGE]:
{
    "instruction": """[INPUT]""",
}

Here are the outputs of the models, also
in [LANGUAGE]:
[
    {

Src \ Tgt	De	En	Es	Ru	Tr	Vi
De	52.3	50.8	63.0	66.7	63.0	60.4
En	56.4	55.5	66.1	70.7	67.2	63.1
Es	51.9	51.2	62.4	66.0	64.4	57.5
Ru	48.1	46.5	59.2	63.6	59.0	56.3
Tr	53.3	52.9	62.6	66.6	60.4	59.0
Vi	46.5	48.2	60.0	65.6	62.1	58.0

Table 6: Cross-lingual alignment results using **best-of-n** with $n = 64$, for the **summarization** task, measured in win rate (%) against the target-language SFT model as judged by **PaLM-2-L** (Figure 4).

Src \ Tgt	En	Es	Ru
En	62.9	65.0	59.6
Es	59.1	62.4	57.6
Ru	53.4	54.3	52.5

Table 7: Cross-lingual alignment results using **best-of-n** with $n = 64$, for the **dialog generation** task, measured in win rate (%) against the target-language SFT model as judged by **PaLM-2-L** (Figure 4).

```
        "model": "model_1",
        "answer": """[GENERATION1]"""
    },
    {
        "model": "model_2",
        "answer": """[GENERATION2]"""
    }
]
```

Respond 1 or 2 to indicate the better
output. Please provide the ranking that
the majority of humans would give.

Better output=

H Raw Results

We show the raw numerical results that correspond
to our plots in Table 6 to 25.

1350

Src \ Tgt	De	En	Es	Ru	Tr	Vi
De	59.4	61.0	59.4	49.6	52.5	59.3
En	55.9	59.9	58.5	52.6	54.8	56.6
Es	52.0	56.1	56.8	53.0	55.0	49.9
Ru	54.8	55.2	56.8	51.8	53.3	52.2
Tr	53.1	54.6	55.7	53.1	53.4	56.3
Vi	63.9	61.8	65.2	54.6	55.1	53.6

Table 8: Cross-lingual alignment results using **RL**, for the **summarization** task, measured in win rate (%) against the target-language SFT model as judged by **PaLM-2-L** (Figure 4).

Src \ Tgt	De	En	Es	Ru	Tr	Vi
De	59.8	59.9	58.4	50.0	55.8	62.4
En	59.4	61.8	59.7	52.1	59.6	61.2
Es	57.6	59.7	58.8	52.0	60.4	60.1
Ru	56.9	56.5	56.4	52.0	57.4	58.0
Tr	59.9	60.7	59.0	52.2	60.1	62.8
Vi	60.5	64.1	63.1	52.5	64.4	61.6

Table 12: Cross-lingual alignment results using **RL**, for the **summarization** task, measured in win rate (%) against the target-language SFT model as judged by **GPT-4** (Figure 9).

Src \ Tgt	En	Es	Ru
En	53.1	54.5	53.5
Es	49.9	51.1	47.5
Ru	51.2	52.7	52.5

Table 9: Cross-lingual alignment results using **RL**, for the **dialog generation** task, measured in win rate (%) against the target-language SFT model as judged by **PaLM-2-L** (Figure 4).

Src \ Tgt	En	Es	Ru
En	51.7	51.9	51.5
Es	49.9	51.5	52.5
Ru	48.5	51.6	51.5

Table 13: Cross-lingual alignment results using **RL**, for the **dialog generation** task, measured in win rate (%) against the target-language SFT model as judged by **GPT-4** (Figure 9).

Src \ Tgt	De	En	Es	Ru	Tr	Vi
De	49.0	50.2	58.2	63.6	57.6	56.6
En	52.6	56.6	62.7	70.2	67.0	62.1
Es	51.7	54.1	59.8	65.9	63.6	59.2
Ru	48.7	51.2	56.0	63.0	59.0	56.8
Tr	56.7	57.8	62.3	69.5	66.6	61.5
Vi	45.2	52.1	56.6	62.8	60.5	56.5

Table 10: Cross-lingual alignment results using **best-of-n** with $n = 64$, for the **summarization** task, measured in win rate (%) against the target-language SFT model as judged by **GPT-4** (Figure 9).

Src \ Tgt	En	Es
De	61.0	64.0
En	60.9	67.4
Es	62.6	69.0
Ru	51.9	63.4
Tr	61.8	66.3
Vi	52.3	61.2

Table 14: Cross-lingual alignment results using **best-of-n**, for the **summarization** task, measured in win rate (%) against the target-language SFT model as judged by **human evaluators** (Figure 2).

Src \ Tgt	En	Es	Ru
En	53.7	58.0	60.6
Es	50.7	56.6	56.6
Ru	50.4	48.6	48.5

Table 11: Cross-lingual alignment results using **best-of-n** with $n = 64$, for the **dialog generation** task, measured in win rate (%) against the target-language SFT model as judged by **GPT-4** (Figure 9).

Src \ Tgt	En	Es
De	64.4	64.2
En	61.4	65.9
Es	58.7	62.7
Ru	61.9	60.6
Tr	63.3	64.9
Vi	66.2	64.7

Table 15: Cross-lingual alignment results using **RL**, for the **summarization** task, measured in win rate (%) against the target-language SFT model as judged by **human evaluators** (Figure 2).

Src \ Tgt	En	Es
En	67.6	52.0
Es	71.4	56.4

Table 16: Cross-lingual alignment results using **best-of-n** with $n = 64$, for the **dialog generation** task, measured in win rate (%) against the target-language SFT model as judged by **human evaluators** (Figure 2).

Src \ Tgt	De	En	Es	Ru	Tr	Vi
De	–	50.0	61.9	66.1	66.1	54.6
En	47.9	–	63.3	64.9	64.5	53.1
Es	50.6	52.9	–	64.1	64.5	59.0
Ru	47.4	51.2	60.3	–	63.3	57.7
Tr	50.6	52.5	61.8	65.6	–	50.8
Vi	42.0	50.8	59.1	64.4	63.6	–

Table 17: Alignment quality using RM trained by translating the source language data into the target language using best-of-n with $n = 64$, for the summarization task, measured in win rate (%) against the target-language SFT model as judged by PaLM-2-L (§5.1).

Src \ Tgt	De	En	Es	Ru	Tr	Vi
De	71.0	64.8	68.0	67.9	67.5	67.7
En	62.2	67.4	67.9	66.3	66.5	70.8
Es	67.4	62.7	72.3	69.7	71.4	65.2
Ru	66.5	61.3	65.4	65.7	66.5	63.6
Tr	66.8	64.6	68.5	69.1	73.2	68.7
Vi	63.0	66.7	68.6	66.5	67.8	71.3
Majority	52.9	59.5	63.1	55.1	56.2	67.9
Length	56.6	59.5	63.1	55.1	55.2	67.9

Table 18: RM generalizability within the reward modeling task evaluated by accuracy (%) on in-task validation data for the summarization task, on the six Seahorse languages, as well as the majority baseline and the length baseline (§6.1) (Figure 7).

Src \ Tgt	En	Es	Ru
En	68.4	68.4	76.3
Es	65.4	67.8	77.0
Ru	56.6	63.5	64.4
Length	66.1	68.1	71.1

Table 19: RM generalizability within the reward modeling task evaluated by accuracy (%) on in-task validation data for the dialog generation task, in three languages, as well as the length baseline (§6.1) (Figure 7).

Src \ Tgt	De	En	Es	Ru	Tr	Vi
De	0.92	0.78	0.83	0.01	0.37	1.92
En	1.50	1.32	1.01	0.02	0.83	3.30
Es	1.78	1.63	1.51	0.10	1.39	3.92
Ru	0.79	0.45	0.46	0.02	0.36	1.26
Tr	2.20	1.91	1.83	0.15	1.34	4.28
Vi	1.78	2.52	1.74	0.02	1.47	4.37

Table 20: KL-divergence of the RL models from the corresponding target-language SFT model for the summarization task (Figure 8).

Lg.	De	En	Es	Ru	Tr	Vi
Mono.	36.2	38.9	32.9	16.9	35.2	41.8
Lg→En	27.8	–	27.1	22.4	28.2	26.7
En→Lg	16.1	–	24.6	13.6	29.9	40.3
En→Lg→En	36.5	–	36.1	35.4	36.5	35.8
Lg→En→Lg	32.5	–	26.6	12.2	32.1	34.9

Table 21: ROUGE-L score when the SFT model is trained using different strategies, either monolingually, translated from a source language, or back-translated into a source language and then back (Figure 5(a)).

Lg.	De	Es	Ru	Tr	Vi
Target-language SFT; RM transfer only					
Lg→En	50.8	51.2	46.5	52.9	48.2
En→Lg	56.4	66.1	70.7	67.2	63.1
(Back-)Translated SFT					
Lg→En	36.6	26.6	29.8	37.5	31.8
En→Lg	14.4	43.5	43.9	47.1	41.6
En→Lg→En	42.7	43.2	40.1	41.4	37.1
Lg→En→Lg	45.3	54.0	60.1	61.7	51.1

Table 22: Alignment performance using best-of-n, measured in the win rate against the monolingual target language SFT model as judged by PaLM-2-L, when the SFT model is trained using different strategies. The first section uses a SFT model that is trained on target-language datasets (same as Table 6), while the second uses translated or back-translated SFT data (Figure 5(b)).

Lg.	De	Es	Ru	Tr	Vi
Target-language SFT; RM transfer only					
Lg→En	38.3	32.4	38.6	32.9	29.2
En→Lg	62.8	59.4	53.7	47.4	66.4
(Back-)Translated SFT					
Lg→En	40.5	29.1	33.2	26.0	19.4
En→Lg	45.7	50.3	60.3	37.1	67.6
En→Lg→En	31.4	33.9	34.0	40.8	31.7
Lg→En→Lg	40.3	31.2	40.1	45.9	61.4

Table 23: Alignment performance using best-of-n, measured in the win rate against the monolingual target language SFT model as judged by PaLM-2-L, when the SFT model is trained using different strategies. The first section uses a SFT model that is trained on target-language datasets, while the second uses translated or back-translated SFT data. Here, we only consider the WikiLingua dataset for both SFT and RM (Figure 5(c)).

Lg.	De	Es	Ru	Tr	Vi
Target-language SFT; RM transfer only					
Lg→En	61.0	56.1	55.2	54.6	61.8
En→Lg	55.9	58.5	52.6	54.8	56.6
(Back-)Translated SFT					
Lg→En	60.2	37.5	22.7	54.9	19.2
En→Lg	28.8	57.0	56.5	59.6	51.9
En→Lg→En	47.5	46.7	42.1	42.4	48.3
Lg→En→Lg	44.7	45.1	46.6	49.5	30.7

Table 24: Alignment performance using RL, measured in the win rate against the monolingual target language SFT model as judged by PaLM-2-L, when the SFT model is trained using different strategies. The first section uses a SFT model that is trained on target-language datasets, while the second uses translated or back-translated SFT data (Figure 5(d)).

Src \ Tgt	De	En	Es	Ru	Tr	Vi
De	52.3	50.8	63.0	66.7	63.0	60.4
En	56.4	55.5	66.1	70.7	67.2	63.1
De + En	56.6	55.7	66.6	70.6	66.7	64.1

Table 25: Alignment performance using best-of-n, measured in the win rate against the monolingual target language SFT model as judged by PaLM-2-L, when using either a monolingual RM (same as Table 6) or a bilingual RM (Figure 11).