Cross-lingual Transfer of Reward Models in Multilingual Alignment

Jiwoo Hong[*†] **Noah Lee**[*†] **Rodrigo Martínez-Castaño**[§]
César Rodríguez[§] **James Thorne**[†]

[†]KAIST AI [§]IQ.WIKI

[†]{jiwoo_hong,noah.lee,thorne}@kaist.ac.kr
[§]{rodrigo,cesar}@iq.wiki

Abstract

Reinforcement learning with human feedback (RLHF) is shown to largely benefit from precise reward models (RMs). However, recent studies in reward modeling schemes are skewed towards English, limiting the applicability of RLHF in multilingual alignments. In this work, we investigate the cross-lingual transfer of RMs trained in diverse languages, primarily from English. Our experimental results demonstrate the strong cross-lingual transfer of English RMs, exceeding target language RMs by 3-4% average increase in `Mutlilingual RewardBench`. Furthermore, we analyze the cross-lingual transfer of RMs through the representation shifts. Finally, we perform multilingual alignment to exemplify how cross-lingual transfer in RM propagates to enhanced multilingual instruction-following capability, along with extensive analyses on off-the-shelf RMs. We release the code,[1] model and data.[2]

1 Introduction

Recent advances in reinforcement learning with human feedback (RLHF) as a large language model (LLM) post-training technique (Christiano et al., 2017; Ziegler et al., 2020) highlight the importance of having high-quality data (Wang et al., 2024f; Dubey et al., 2024) and reward model (RM) (Ethayarajh et al., 2022; Gao et al., 2023; Ji et al., 2023; Wang et al., 2024a,e). Leveraging synthetic data has contributed to building stronger English RMs due to their efficiency and scalability (Cui et al., 2024; Wang et al., 2024b; Zhu et al., 2024).

Nevertheless, adopting RMs for non-English languages is heavily understudied. While LLM-as-a-Judge can be used as a generative reward model for multilingual RLHF settings (Son et al., 2024), generative RMs have been shown to underperform traditional RMs (Lambert et al., 2024; Wang et al., 2024b). Meanwhile, Wu et al. (2024) empirically demonstrates the possibilities of cross-lingual transfer in RMs, but the findings were limited to simple tasks and encoder-decoder models.

In this paper, we show that RMs trained on English-only datasets (*i.e.,* English RMs) display strong cross-lingual transfer when built on top of multilingual pre-trained language models (MLMs). We first demonstrate the cross-lingual transfer of English RMs by consistently outperforming target language RMs in `Multilingual RewardBench`. Then, we explain it with two reasons: **1)** English preserves representations of the initial MLMs (Section 3.1), and **2)** representations of MLMs inherently have a strong understanding of languages (Section 3.2), concluding that RMs should preserve representations of MLMs for generalizability. Additional analysis of off-the-shelf RMs supports our findings by both classifier and generative RMs based on MLMs having strong cross-lingual transfer. Finally, multilingual alignment experiments exhibit the propagation of strong cross-lingual transfer in English RMs to downstream usage, having an average win rate increase of 9.5% across four non-English languages.

2 English as a *Lingua Franca* in RMs

We empirically verify the cross-lingual transfer in reward models (RMs) trained with different languages, thereby showing that the English preference data is a *lingua franca* in reward modeling.

2.1 Background

Cross-lingual transfer Training multilingual language models (MLMs) at scale has shown to incur *cross-lingual transfer* in both encoder-only (Devlin et al., 2019; Conneau et al., 2020; Chi et al., 2022) and encoder-decoder (Xue et al., 2021) transformer architectures. Recently, studies revealed the

[*]Equal Contribution
[1]Code - IQ-KAIST/rm-lingual-transfer
[2]Data & Models - HF Collection

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 82–94
April 29 - May 4, 2025

RewardBench	Category	Llama-3.2-3B-IT Chat	Chat(H)	Safety	Reason	Avg.	Qwen2.5-3B-IT Chat	Chat(H)	Safety	Reason	Avg.
Spanish	**Target**	79.1	67.3	88.0	65.5	75.0	80.7	68.2	84.8	68.2	75.5
	English	86.3	69.3	89.3	72.4	79.3	82.7	68.0	88.3	73.6	78.1
	Δ	+7.2	+2.0	+1.3	+6.9	+4.3	+2.0	-0.2	+3.5	+5.4	+2.6
Italian	**Target**	75.4	62.5	88.5	65.7	73.0	77.1	67.8	85.7	72.8	75.8
	English	83.0	69.3	88.7	75.1	79.0	83.2	68.2	88.4	76.0	79.0
	Δ	+7.6	+6.8	+0.2	+9.4	+6.0	+6.1	+0.4	+2.7	+3.2	+3.2
Korean	**Target**	69.6	58.8	80.9	60.1	67.3	68.4	63.2	80.9	61.4	68.5
	English	69.8	59.4	84.3	73.0	71.6	70.7	61.6	85.4	73.6	72.8
	Δ	+0.2	+0.6	+3.4	+12.9	+4.3	+2.3	-1.6	+4.5	+12.2	+4.3
Chinese	**Target**	68.7	59.9	81.2	52.6	65.6	69.8	64.7	81.8	61.3	69.4
	English	54.7	64.0	82.6	79.3	70.2	58.7	67.8	84.3	78.2	72.2
	Δ	-14.0	+4.1	+1.4	+26.7	+4.6	-11.1	+3.1	+2.5	+16.9	+2.8

Table 1: `Multilingual RewardBench` evaluation results on the target language ("Target") and English ("English") RMs. "Δ" denotes the accuracy gain of English RMs compared to the target language RMs. English RMs show higher average scores in the lingual axis than target language RMs. Also, English RMs excel target language RMs in reasoning ("Reason") tasks with diverse evaluation sub-categories.

implications of cross-lingual transfer in decoder-only models as well (Üstün et al., 2024; Wang et al., 2024c); however, they were limited to generative tasks (Zhang et al., 2024) or downstream alignment-tuning only (Dang et al., 2024).

Reward modeling Reward models are trained as a classifier (Christiano et al., 2017) to return a scalar value $r_\theta(\cdot)$ with the objective with the Bradley-Terry model (Bradley and Terry, 1952):

$$\mathcal{L}_{\text{RM}} = \sigma\left(r_\theta(x, y_w) - r_\theta(x, y_l)\right),$$

with the prompt x and corresponding preferred and dispreferred responses y_w and y_l. While crucial in alignment-tuning (Rafailov et al., 2024; Hong et al., 2024; Meng et al., 2024), reward modeling schemes for multilingual usage are still understudied. Motivated by this research opportunity, we study the cross-lingual transfer of English-focused RMs with recent autoregressive models and how it propagates to downstream multilingual alignment.

2.2 Experimental Details

Dataset We curate a synthetic preference dataset of 86k instances[3] from five representative English preference datasets: SafeRLHF (Dai et al., 2024), WildGuard (Han et al., 2024), HelpSteer2 (Wang et al., 2024e), Offsetbias (Park et al., 2024), and Magpie (Xu et al., 2024b). Using English data, we create four parallel machine-translated versions[4], utilizing X-ALMA (Xu et al., 2024a).

Models Two state-of-the-art 3B multilingual pre-trained language models are fine-tuned[5] as reward models: Llama-3.2-3B-Instruct (Dubey et al., 2024) and Qwen2.5-3B-Instruct (Yang et al., 2024).

Evaluation We prepare four non-English `Multilingual RewardBench` by translating `RewardBench` (Lambert et al., 2024) to assess the cross-lingual transfer in RMs.

2.3 Results and Analysis

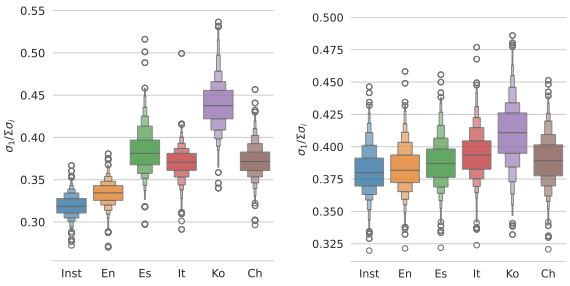
English RMs show strongest cross-lingual transfer Average reward model accuracy ("Avg") in Table 1 shows that English RMs surpass target language RMs in general. Specifically, Llama-3.2-3B gained at least 4.3%, where the cross-lingual generalizability of English RMs is more highlighted than Qwen2.5-3B, which gained at most 4.3%. However, considering that all Qwen-based target language RMs outperform the Llama-based target language RMs, Qwen2.5-3B is shown to be a better model choice for training a language-specific RM.

Reasoning tasks significantly benefit from cross-lingual transfer Generalizability of English RMs is best highlighted in the reasoning tasks ("Reason") in Table 1, especially in non-Latin languages. Non-Latin languages, Korean and Chinese, improved significantly in English RMs compared to target language RMs, exceeding 12% and 27% in Chinese, for instance.

[3] Refer to Appendix A for detailed process.

[4] Spanish (Sp), Italian (It), Korean (Ko), and Chinese (Ch)

[5] Refer to Appendix B for detailed hyperparameters.



(a) Llama-3.2-3B (b) Qwen2.5-3B

Figure 1: Proportion of the largest singular value in the concatenated hidden states for fixed context translated in five languages with RMs trained in each language. While English ("En") best preserves the representation diversity of the base model ("Inst'), Korean ("Ko") leads to the most homogeneous representations.

3 Analysis on Lingual Tranfer of MLM

This section provides empirical and theoretical insights on *why* English is *lingua franca* in reward modeling, given a multilingual language model (MLM) using two arguments: **1)** English acts as a *lingua franca* in reward modeling because it best preserves the representations of the base model, and **2)** representations in MLMs *should* be preserved since they are inherently effective in language-aware encoding.

3.1 English preserves general representations

Non-English reward modeling is detrimental to generalizability In general, the generalizability of the downstream model is closely connected to *how much the representations are preserved* during the fine-tuning (Aghajanyan et al., 2021; Razdaibiedina et al., 2023). We demonstrate this in RMs by ablating over different languages and tasks. We assess the general representation preservation of RMs used in Section 2 by comparing their hidden states against the initial model. To do so, we measure how much the distinct representations are collapsed into similar spaces in Figure 1. In specific, we construct a matrix of the last hidden states $\mathcal{H}_\theta(x) \in \mathbb{R}^{5\times d_{\text{model}}}$ across five languages using multilingual dataset BeleBele (Bandarkar et al., 2024):

$$\mathcal{H}_\theta(x) = \text{concat}\left[\left\{H_\theta^l(x_l)\right\}_{l\in L}\right] \in \mathbb{R}^{|L|\times d_{\text{model}}},$$

where $H_\theta^l(x) \in \mathbb{R}^{d_{\text{model}}}$ refers to the last hidden state of the model θ for sequence x_l in the language l, but with fixed context. Then, we measure the proportion of the largest singular value in $\mathcal{H}_\theta(x)$:

$$f_\theta(x) = \frac{\sigma_1}{\sum_{i=1}^{|L|}\sigma_i}, S = \text{diag}\left(\sigma_1, \ldots, \sigma_{|L|}\right),$$

with S as singular value matrix of $\mathcal{H}_\theta(x)$. Intuitively, having $f_\theta(x)$ close to 1 imples the hidden states in different languages are homogeneous: *i.e.,* representations are embedded into similar space.

In Figure 1, we plot $f_\theta(x)$ with different RMs. English RMs best preserve the representations by staying close to the base instruct model ("Inst"). On the other hand, Korean RMs ("Ko") tend to deviate the most from the base model, thereby homogenizing the multilingual representations the most. Both observations were more extreme in Llama-3.2-3B.

General representation preservation is crucial for cross-lingual/task transfer Notably, the proclivity in general representation preservation in Figure 1 aligns with the accuracy in Table 1. Non-English RMs with Llama-3.2-3B tend to introduce stronger representation collapse than Qwen2.5-3B in Figure 1. This aligns with Section 2.3 as Llama-3.2-3B gets more severe degradation using target language RMs, implying the significance of representation preservation in cross-lingual transfer.

Furthermore, the same tendency holds for cross-task analysis. `RewardBench` has especially fine-grained divisions under the reasoning category (*e.g.,* Java, Python, Rust, math) compared to other categories. Thus, strong generalization abilities are crucial to achieving decent scores in the reasoning category. Interestingly, English RMs dominate other languages in reasoning despite the fixed data across the languages in Table 1, which strongly supports the significance of representation preservation in cross-task generalization.

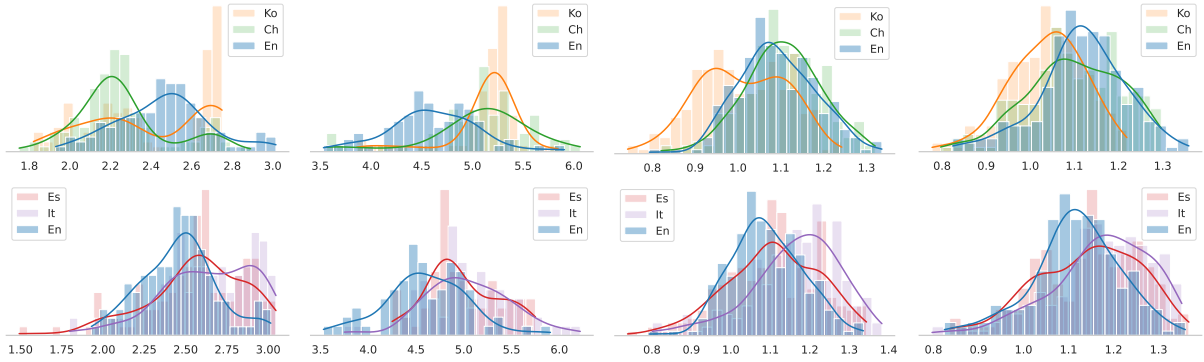
3.2 MLM representations are language-aware

In autoregressive language models (Radford et al., 2019) with tied embeddings (Jiang et al., 2023; Team, 2024a), the logits for next token is:

$$h_t \cdot E = \left[\|h_t\| \cdot \|e_i\| \cdot \cos\left(\theta_i\right)\right]_{i=1}^{|V|},$$

where θ_i is the angle between h_t and e_i. Therefore, the capability of language models in generative tasks is closely related to having *good representations* (Edunov et al., 2019) that could accurately align with the ideal next token.

Token embeddings are a good proxy to understand the effectiveness of representations as they



(a) OLMo-1B (b) SmolLM-1.7B (c) Qwen2.5-3B (d) Llama-3.2-3B

Figure 2: Embedding norm distribution comparison between English and four other languages (2 non-Latin (top), 2 Latin (bottom)) across four language models: OLMo-1B and SmolLM-1.7B (monolingual pre-training) and Qwen2.5-3B and Llama-3.2-3B (multilingual pre-training). While English and non-English token embedding norm distributions of OLMo-1B and SmolLM-1.7B are distinct, they are similar in Qwen2.5-3B and Llama-3.2-3B.

imply the imbalance in pre-training corpora (Chung et al., 2024), especially by *linguality* in this study (Wen-Yi and Mimno, 2023). Thus, we can infer that language models with similar embedding norm distribution across the language will have decoder layers that can return language-aware fine-grained hidden states, which deserve to be preserved for their generalizability.

MLMs have similar token embedding norm distributions across the language We validate this point by comparing the two models in Section 2 with two monolingual pre-trained language models: OLMo-1B (Groeneveld et al., 2024) and SmolLM-1.7B (Allal et al., 2024). We clarify the lingualities in each model's pre-training in Appendix C.

We collect the disjoint language-specific token embedding norms for each model:

$$\mathbf{e}_l = \{\|e_j\|\}_{j \in A_l}, A_l \subset V, \bigcap_{l \in L} A_l = \varnothing$$

where A_l is the token indices of language l in V. We compare $\mathbf{e}_L$ distribution over five languages.

In Figure 2, the distribution for English in SmolLM-1.7B and OLMo-1B are distinct from four languages, especially Korean and Chinese, which are non-Latin languages that do not share similar alphabets. However, Qwen2.5-3B and Llama-3.2-3B have similar ranges and distributions across the languages, even in non-Latin languages.

Thus, we can infer that Qwen2.5-3B and Llama-3.2-3B, as MLMs, are sufficiently trained on the multilingual corpus to encode information with diverse linguality by having similar embedding norm distributions across the languages (Dagan et al., 2024; Chung et al., 2024). This supports why representation preservation is a crucial condition for generalizable RMs with MLMs, as discussed in Section 3.1.

4 Multilingual Alignment using RM

In this section, we perform experiments to outline the effects of using the reward models (RMs) from Section 2 and how their cross-lingual transfer can propagate to the actual alignment process.

4.1 Experimental Details

We sample 10k prompts from the cleaned UltraFeedback dataset (Bartolome et al., 2023; Cui et al., 2024) and translate prompts across target languages. Then, we sample four responses per prompt with Qwen2.5-7B-Instruct (Team, 2024b) and label them with desired RMs. By selecting the responses with the highest and lowest rewards, we prepare pairwise preference data. We train Qwen2.5-7B-Instruct on each language from the newly curated datasets with Direct Preference Optimization (Rafailov et al., 2024, DPO). Refer to Appendix B for the detailed setup.

Evaluation We evaluate the trained model's language-specific instruction-following capability with `Multilingual AlpacaEval`, adopted from the instances and evaluation pipeline of `AlpacaEval` (Li et al., 2023). We report the detailed process and configurations in Appendix D.

4.2 Results and Analysis

English RM largely improves base models in *every* language As shown in Figure 3, models

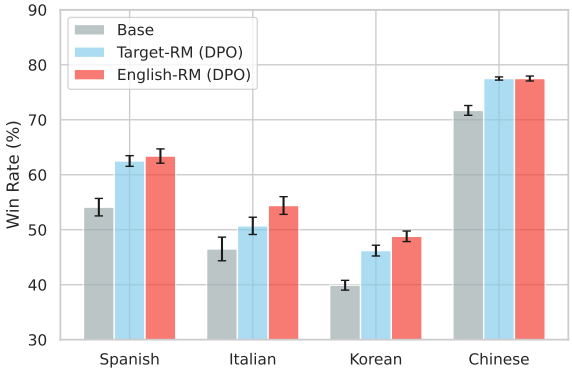


Figure 3: `Multilingual AlpacaEval` results of Qwen2.5-7B-Instruct models fine-tuned with DPO on on-policy generations for four non-English languages over fine runs. The alignment data were labeled with either English RM or target language RM. Results are averaged over 5 runs.

aligned with English RM show a notable leap compared to Qwen2.5-7B-Instruct ("Base"), by increasing up to 9.3% point in Spanish. As the win rate was measured against GPT-4-Turbo, a strong proprietary language model, such enhancements strongly support the validity of using English RMs for multilingual alignment in desired languages.

Exploiting English RMs is a desirable choice in multilingual alignment We emphasize that using high-quality English preference data of better accessibility is a decent choice, considering the efficiency and efficacy in real-world cases. In Figure 3, models aligned with English RM outperformed or at least on par with ones with target language RMs, tied only in Chinese. Thus, adopting an English RM for multilingual alignment is a cost-efficient yet performant alternative, discarding the need for scaled translations for the reward model.

5 Cross-lingual Transfer of External RMs

Along with the controlled comparisons in Section 2, we analyze the cross-lingual transfer in off-the-shelf models on the original `RewardBench` through `Multilingual RewardBench`. To ensure diversity in reward modeling schemes, we selected two classifier reward models (RMs), ArmoRM-8B (Wang et al., 2024b) and OffsetBias-8B (Park et al., 2024), alongside two generative RMs, GPT-4o[6] and Self-Taught-Llama-70B (Wang et al., 2024d).

[6] https://platform.openai.com/docs/models/gpt-4o

Model	En	Es	It	Ko	Ch
ArmoRM-8B	**90.4**	80.1	78.9	71.5	69.6
OffsetBias-8B	89.4	78.9	79.5	74.5	73.1
GPT-4o†	86.7	80.4	78.6	75.2	72.1
ST-L-70B*	90.0	**83.1**	**81.5**	**75.6**	**74.1**

Table 2: Averaged Multilingual RewardBench results in two classifier RMs (top) and two generative RMs (bottom). Off-the-shelf RMs based on MLMs show strong cross-lingual transfer as in Table 1.

Classifier RMs Two classifier RMs are both trained on top of Llama3-8B-Instruct (Dubey et al., 2024), which are based on multilingual pre-trained language models (MLMs) as discussed in Appendix C. As in Table 1, these RMs also demonstrate strong cross-lingual transfer in four languages, mostly exceeding 70% accuracy across the board in Table 2.

Generative reward models Interestingly, we can observe strong cross-lingual transfer in the generative RMs in Table 2, as in the classifier RMs. As discussed in Section 3.2, fine-grained representation learning is a crucial component for having strong downstream generative abilities. While the extent of multilingual pre-training in GPT-4o is not verifiable, GPT-4o has the least decrement in non-English settings. Meantime, Self-Taught-Llama-70B with extensive multilingual pre-training demonstrates the strongest cross-lingual transfer, achieving the best accuracies in all four non-English `Multilingual RewardBench`.

Conclusion

We empirically demonstrate English as a *lingua franca* in reward modeling, given recent multilingual pre-trained language models (MLMs). We explain this with two consecutive arguments. First, English reward models (RMs) best preserve the representations of initial MLMs, while other languages induce representation collapse. Second, MLM representations inherently have a rich understanding of languages and tasks, making them valuable to preserve in downstream tasks. By extending our analysis to the off-the-shelf reward models, we show that using MLMs for reward modeling is crucial for eliciting strong cross-lingual transfer. Through strong cross-lingual transfer in English RMs, we establish a concrete foundation for exploiting English RMs for multilingual alignment.

Limitations

To extend to more languages and evaluation benchmarks, we have mainly utilized a 3B LLM to train the reward model (RM) with only 86k instances. However, as outlined in Appendix E, the 3B RMs are on par with a state-of-the-art RM, ArmoRM, which was trained with over 550k instances. Future works on the effects of data size and mixture will provide an enhanced understanding of our work.

Also, in Section 4, we use the AlpacaEval evaluation setup, which utilizes LLM-generated reference responses and LLM-as-a-Judge to select a winning response. Therefore, while we show a vast increase in post-training alignment, the process relies on the multilinguality of OpenAI models and the evaluation biases of the LLM-based evaluations outlined in Zheng et al., 2023.

Acknowledgment

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Technology research to ensure authenticity and consistency of results generated by AI) and (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)).

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2021. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Leandro von Werra, and Thomas Wolf. 2024. Smollm - blazingly fast and remarkably powerful.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Alvaro Bartolome, Gabriel Martin, and Daniel Vila. 2023. Notus. https://github.com/argilla-io/notus.

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.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Bo Zheng, Saksham Singhal, Payal Bajaj, Xia Song, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2022. XLM-E: Cross-lingual language model pre-training via ELECTRA. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6170–6182, Dublin, Ireland. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Woojin Chung, Jiwoo Hong, Na Min An, James Thorne, and Se-Young Yun. 2024. Stable language model pre-training by reducing embedding variability. *Preprint*, arXiv:2409.07787.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. 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.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: Boosting language models with scaled ai feedback. *Preprint*, arXiv:2310.01377.

Gautier Dagan, Gabriel Synnaeve, and Baptiste Roziere. 2024. Getting the most out of your tokenizer for pre-training and domain adaptation. In *Forty-first International Conference on Machine Learning*.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. 2024. Safe rlhf: Safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations*.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. Rlhf can speak many languages: Unlocking multilingual preference optimization for llms. *arXiv preprint arXiv:2407.02552*.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. *9th International Conference on Learning Representations, ICLR*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information*

Processing Systems, volume 36, pages 10088–10115. Curran Associates, Inc.

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.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and Aurelien Rodriguez et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Sergey Edunov, Alexei Baevski, and Michael Auli. 2019. Pre-trained language model representations for language generation. 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 4052–4059, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. 2022. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. 2024. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Preprint*, arXiv:2406.18495.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *EMNLP*.

Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, and Siyu Zhu. 2024. Liger-kernel: Efficient triton kernels for llm training.

Shengyi Huang, Michael Noukhovitch, Arian Hosseini, Kashif Rasul, Weixun Wang, and Lewis Tunstall. 2024. The n+ implementation details of RLHF with PPO: A case study on TL;DR summarization. In *First Conference on Language Modeling*.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Raghavi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Rewardbench: Evaluating reward models for language modeling. *CoRR*, abs/2403.13787.

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

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. 2024. Offsetbias: Leveraging debiased data for tuning evaluators. *Preprint*, arXiv:2407.06551.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2406.17557.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20. IEEE Press.

Anastasia Razdaibiedina, Ashish Khetan, Zohar Karnin, Daniel Khashabi, and Vivek Madan. 2023. Representation projection invariance mitigates representation collapse. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14638–14664, Singapore. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Guijin Son, Hyunwoo Ko, Hoyoung Lee, Yewon Kim, and Seunghyeok Hong. 2024. Llm-as-a-judge & reward model: What they can and cannot do. *arXiv preprint arXiv:2409.11239*.

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.

Gemma Team. 2024a. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Qwen Team. 2024b. Qwen2.5: A party of foundation models.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, Songyang Gao, Nuo Xu, Yuhao Zhou, Xiaoran Fan, Zhiheng Xi, Jun Zhao, Xiao Wang, Tao Ji, Hang Yan, Lixing Shen, Zhan Chen, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. 2024a. Secrets of rlhf in large language models part ii: Reward modeling. *CoRR*, abs/2401.06080.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024b. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845*.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024c. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12159–12173, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024d. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024e. Helpsteer2: Open-source dataset for training top-performing reward models. *Preprint*, arXiv:2406.08673.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope,

and Oleksii Kuchaiev. 2024f. HelpSteer: Multi-attribute helpfulness dataset for SteerLM. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3371–3384, Mexico City, Mexico. Association for Computational Linguistics.

Andrea W Wen-Yi and David Mimno. 2023. Hyperpolyglot LLMs: Cross-lingual interpretability in token embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1124–1131, Singapore. Association for Computational Linguistics.

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. 2024. Reuse your rewards: Reward model transfer for zero-shot cross-lingual alignment. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2024a. X-alma: Plug & play modules and adaptive rejection for quality translation at scale. *Preprint*, arXiv:2410.03115.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024b. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *Preprint*, arXiv:2406.08464.

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.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024. PLUG: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7025–7046, Bangkok, Thailand. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2024. Starling-7b: Improving helpfulness and harmlessness with RLAIF. In *First Conference on Language Modeling*.

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. *Preprint*, arXiv:1909.08593.

A Data Curation

We used full datasets for HelpSteer2, SafeRLHF, and Offsetbias. We filtered the prompts with one harmful and unharmful response each for WildGuard, finally having 8,383 instances. Lastly, we randomly sample 60,000 instances from the synthetic preference dataset comprising responses from Llama-3-70B-Instruct (Dubey et al., 2024) and Gemma-2-9B-It (Team, 2024a) labeled with ArmoRM (Wang et al., 2024b). From the 108k instances, we finally select 80% of instances as the train set.

B Training Configurations

Both reward modeling and downstream on-policy preference optimization were down using Hugging Face TRL library (von Werra et al., 2020) on 4 NVIDIA A100 GPUs with Accelerate (Gugger et al., 2022) and DeepSpeed ZeRO 3 (Rajbhandari et al., 2020), and Paged AdamW optimizer (Loshchilov and Hutter, 2019; Dettmers et al., 2023) with 8-bit precision (Dettmers et al., 2022).

B.1 Reward Modeling

We used a maximum learning rate of $1e-5$ and 10% of warm-up followed by cosine decay. The projection head for the reward model was initialized with $\mathcal{N}\left(0, 1/\sqrt{d_{\text{model}}+1}\right)$ (Stiennon et al., 2020; Huang et al., 2024). The global batch was set to 128.

B.2 On-Policy Preference Optimization

We fine-tune Qwen2.5-7B-Instruct (Team, 2024b) with DPO using Liger-kernel (Hsu et al., 2024). We use a cosine decaying learning rate scheduler for single epoch training.

DPO configurations We apply $\beta = 0.1$ with the learning rate of $5\text{e}-7$. The global batch size was set to 32 using gradient accumulation steps of 8 with a per-device batch size of 1, which was the maximum number for NVIDIA A100 80GiB.

Data curation To construct the preference pairs for preference optimization, we sample 4 responses from Qwen-2.5-7B-Instruct. Then, we compute the rewards through the reward models and select the response with the highest and lowest reward values as the preference pairs for training the checkpoints through DPO.

C Linguality in Pre-training

Olmo-1B and SmolLM-1.7B are selectively pre-trained on Dolma (Soldaini et al., 2024) and an English-focused subset of FineWeb (Penedo et al., 2024), respectively: *i.e.,* monolingual pre-training. On the other hand, the Qwen2.5 series is pre-trained on more than 7 trillion tokens comprising more than 30 languages (Yang et al., 2024; Team, 2024b): *i.e.,* multilingual pre-training. Similarly, 8% of 15 trillion tokens for pre-training Llama-3 series were multilingual (Dubey et al., 2024).

D MULTILINGUAL ALPACAEVAL Setup

Starting from the 805 translated prompt instances[7] (Zhang et al., 2024), we compute the language-specific win-rate of the model evaluated by GPT-4o[8] against the reference responses from GPT-4-Turbo[9]. Given the generations from the reference model and aligned model, we adopt a LLM-as-a-Judge evaluation given the evaluation template[10].

[7] https://huggingface.co/datasets/zhihz0535/X-AlpacaEval
[8] https://platform.openai.com/docs/models/gpt-4o
[9] https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4
[10] https://github.com/tatsu-lab/alpaca_eval/blob/main/src/alpaca_eval/evaluators_configs/

E REWARDBENCH Evaluation Results Across Languages

REWARD MODEL	CHAT	CHAT(H)	SAFETY	REASON	AVG.
ARMORM-L3-8B*	96.9	76.8	90.5	97.3	90.4
L32-3B-IT-EN	92.5	81.8	90.2	95.5	90.0
L32-3B-IT-SP	82.1	71.7	88.2	81.5	80.9
L32-3B-IT-IT	86.3	66.0	88.4	75.4	79.0
L32-3B-IT-KO	84.4	70.6	84.8	78.7	79.6
L32-3B-IT-CH	82.4	69.7	85.5	86.6	81.0
Q25-3B-IT-EN	89.1	75.2	87.3	95.4	86.8
Q25-3B-IT-SP	89.7	70.4	85.1	83.2	82.1
Q25-3B-IT-IT	88.3	68.9	86.2	88.8	83.0
Q25-3B-IT-KO	86.3	69.5	84.6	76.8	79.3
Q25-3B-IT-CH	84.6	68.2	84.8	89.1	81.7
Q25-7B-IT-EN	91.3	81.6	90.3	96.5	89.9
Q25-7B-IT-SP	90.5	75.9	89.5	94.1	87.5
Q25-7B-IT-IT	90.8	74.1	88.5	92.5	86.5
Q25-7B-IT-KO	89.4	70.8	87.9	94.9	85.8
Q25-7B-IT-CH	83.2	72.6	87.2	90.8	83.5

Table 3: REWARDBENCH results for reward model comparison across four different categories. (* denotes off-the-shelf models)

REWARD MODEL	CHAT	CHAT(H)	SAFETY	REASON	AVG.
ARMORM-L3-8B*	89.4	64.5	89.0	77.5	80.1
L32-3B-IT-EN	86.3	69.3	89.3	72.4	79.3
L32-3B-IT-SP	79.1	67.3	88.0	65.5	75.0
L32-3B-IT-IT	80.4	63.2	88.0	64.8	74.1
L32-3B-IT-KO	79.1	63.8	84.0	54.8	70.4
L32-3B-IT-CH	77.9	64.9	84.1	59.4	71.6
Q25-3B-IT-EN	82.7	68.0	88.3	73.6	78.1
Q25-3B-IT-SP	80.7	68.2	84.8	68.2	75.5
Q25-3B-IT-IT	78.2	67.5	87.0	73.4	76.6
Q25-3B-IT-KO	77.1	67.1	85.3	58.4	72.0
Q25-3B-IT-CH	78.8	64.5	85.3	76.4	76.2
Q25-7B-IT-EN	82.1	73.7	91.4	73.3	80.1
Q25-7B-IT-SP	84.1	71.5	89.9	78.4	81.0
Q25-7B-IT-IT	84.6	70.0	89.2	78.3	80.5
Q25-7B-IT-KO	84.9	65.8	87.0	76.0	78.4
Q25-7B-IT-CH	83.5	66.0	87.2	69.5	76.5

Table 4: Spanish REWARDBENCH results for reward model comparison across four different categories. (* denotes off-the-shelf models)

Reward Model	Chat	Chat(H)	Safety	Reason	Avg.
ArmoRM-L3-8B*	83.2	65.4	88.6	78.5	78.9
L32-3B-IT-En	83.0	69.3	88.7	75.1	79.0
L32-3B-IT-Sp	74.9	67.8	87.6	65.7	74.0
L32-3B-IT-It	75.4	62.5	88.5	65.7	73.0
L32-3B-IT-Ko	77.7	64.9	84.8	57.1	71.1
L32-3B-IT-Ch	75.4	62.5	84.5	61.7	71.0
Q25-3B-IT-En	83.2	68.2	88.4	76.0	79.0
Q25-3B-IT-Sp	81.0	65.8	84.3	70.9	75.5
Q25-3B-IT-It	77.1	67.8	85.7	72.8	75.8
Q25-3B-IT-Ko	78.8	68.0	82.5	61.7	72.7
Q25-3B-IT-Ch	82.1	64.9	83.7	76.7	76.9
Q25-7B-IT-En	82.4	73.0	89.6	75.1	80.0
Q25-7B-IT-Sp	84.6	69.3	89.1	79.8	80.7
Q25-7B-IT-It	80.2	69.7	87.9	78.5	79.1
Q25-7B-IT-Ko	84.1	64.3	85.8	72.7	76.7
Q25-7B-IT-Ch	81.8	65.8	86.5	67.9	75.5

Table 5: Italian RewardBench results for reward model comparison across four different categories. (* denotes off-the-shelf models)

Reward Model	Chat	Chat(H)	Safety	Reason	Avg.
ArmoRM-L3-8B*	66.5	60.3	83.8	75.3	71.5
L32-3B-IT-En	69.8	59.4	84.3	73.0	71.6
L32-3B-IT-Sp	70.7	60.3	84.0	67.8	70.7
L32-3B-IT-It	74.9	56.6	83.6	66.2	70.3
L32-3B-IT-Ko	69.6	58.8	80.9	60.1	67.3
L32-3B-IT-Ch	69.3	58.3	79.7	59.3	66.7
Q25-3B-IT-En	70.7	61.6	85.4	73.6	72.8
Q25-3B-IT-Sp	74.9	59.6	82.3	69.2	71.5
Q25-3B-IT-It	74.3	62.1	82.0	69.4	71.9
Q25-3B-IT-Ko	68.4	63.2	80.9	61.4	68.5
Q25-3B-IT-Ch	74.3	61.2	82.2	66.2	71.0
Q25-7B-IT-En	68.2	66.2	87.9	70.9	73.3
Q25-7B-IT-Sp	75.7	59.9	86.1	70.4	73.0
Q25-7B-IT-It	76.3	61.0	84.9	68.8	72.7
Q25-7B-IT-Ko	72.9	65.4	84.8	67.6	72.7
Q25-7B-IT-Ch	76.3	63.2	84.6	65.1	72.3

Table 6: Korean RewardBench results for reward model comparison across four different categories. (* denotes off-the-shelf models)

Reward Model	Chat	Chat(H)	Safety	Reason	Avg.
ArmoRM-L3-8B*	60.6	60.5	83.7	73.6	69.6
L32-3B-IT-En	54.7	64.0	82.6	79.3	70.2
L32-3B-IT-Sp	61.2	60.5	82.9	70.5	68.8
L32-3B-IT-It	66.8	57.0	84.9	66.4	68.8
L32-3B-IT-Ko	68.4	61.0	81.1	61.3	67.9
L32-3B-IT-Ch	68.7	59.9	81.2	52.6	65.6
Q25-3B-IT-En	58.7	67.8	84.3	78.2	72.2
Q25-3B-IT-Sp	68.7	62.5	79.5	71.0	70.4
Q25-3B-IT-It	69.8	62.3	81.6	70.6	71.1
Q25-3B-IT-Ko	70.1	61.4	79.7	62.3	68.4
Q25-3B-IT-Ch	69.8	64.7	81.8	61.3	69.4
Q25-7B-IT-En	55.0	66.2	85.7	75.8	70.7
Q25-7B-IT-Sp	71.5	63.4	84.9	72.9	73.2
Q25-7B-IT-It	70.9	60.7	85.7	67.6	71.2
Q25-7B-IT-Ko	73.5	60.7	83.9	70.1	72.1
Q25-7B-IT-Ch	67.9	61.6	84.8	64.1	69.6

Table 7: Chinese RewardBench results for reward model comparison across four different categories. (* denotes off-the-shelf models)