

Equitable Alignment of LLMs Across Languages with Regularized Direct Preference Optimization

Thibaud Leteno^{1,2,*}, Antoine Gourru¹, Charlotte Laclau³, and Christophe Gravier¹

¹Université Jean Monnet Saint-Étienne, CNRS, Institut d’Optique Graduate School,
Laboratoire Hubert Curien UMR 5516, F-42023 Saint-Étienne, France

²INSA Lyon, LIRIS - UMR 5205 CNRS, Lyon, France

³Télécom Paris, Institut Polytechnique de Paris, Paris, France

thibaud.leteno@insa-lyon.fr, {antoine.gourru, christophe.gravier}@univ-st-etienne.fr, charlotte.laclau@telecom-paris.fr

Abstract

Aligning large language models with human values is a critical challenge for ensuring public safety. This alignment is commonly achieved through methods such as Direct Preference Optimization and RLAI/RLHF. Despite extensive research, prior studies show that models often fail to maintain consistent alignment performance across languages, resulting in unequal treatment of users. To address this issue, we introduce a regularization term in the Direct Preference Optimization fine-tuning process that penalizes discrepancies in model behavior between languages using a bilingual corpus. Our results demonstrate that this approach effectively reduces the alignment gap across languages following fine-tuning.¹

1 Introduction

As Large Language Models (LLMs) grow in capability and adoption (Bick et al., 2026), aligning them with human values has emerged as a critical research area in Natural Language Processing (NLP). The goal of alignment is to ensure model behavior remains consistent with societal norms and to prevent toxic outputs, even in complex or ambiguous contexts. The advent of generative LLMs such as GPT-3 (Radford et al., 2019) and LLaMA (Touvron et al., 2023) introduces new risks as these models can produce toxic or biased content without malicious prompts, raising ethical issues regarding their deployment and societal impact (Solaiman et al., 2019). Studies such as Bai et al. (2022) and Ouyang et al. (2022) proposed methods to reduce these risks through reinforcement learning from human feedback (RLHF). Another aspect of the alignment effort focuses on guiding models to fol-

low moral norms, which are often implicit and culturally dependent (Ramezani and Xu, 2023). Prior work in this area follows two main directions: first, evaluating whether LLMs align with specific moral philosophies (Tlaie, 2024; Coleman et al., 2025); and second, assessing model behavior in scenarios requiring moral judgment (Hendrycks et al.; Emelin et al., 2021). To address these challenges, several studies have explored aligning LLMs using Direct Preference Optimization (DPO) (Rafailov et al., 2023), a reward maximization approach that relies on preference data. For example, Li et al. (2024) reduce toxicity in open-ended generations using DPO, while Leteno et al. (2025) demonstrate that LLMs can easily be aligned with moral values using DPO and a limited dataset. However, as noted by Blasi et al. (2022); Lai et al. (2023), the performance of LLMs often varies across languages. The phenomenon also appears in evaluations of moral alignment as observed across languages of different language families in several studies (Hämmerl et al., 2023; Agarwal et al., 2024; Xu et al., 2024).

Related to this work, several studies have explored DPO in multilingual settings. Wu et al. (2024) and Hong et al. (2024) evaluate zero-shot cross-lingual alignment on generative tasks by fine-tuning models with DPO in one language and testing them on another, showing variation of performance across languages. She et al. (2024) and Yang et al. (2025) propose frameworks that leverage English-aligned models to generate preference pairs for multilingual DPO fine-tuning, improving performance across languages. However, alignment disparities between languages, referred to as the alignment gap, remain unaddressed. Furthermore, none of these approaches explicitly aims to close the alignment gap to ensure equitable treatment across languages. On the other hand, Aakanksha et al. (2024) and Li et al. (2024) investigate multilingual alignment in toxic generation.

*Work done while at Université Jean Monnet.

¹Code and data are available anonymously at https://github.com/LetenoThibaud/Equitable_Alignment_Across_Languages

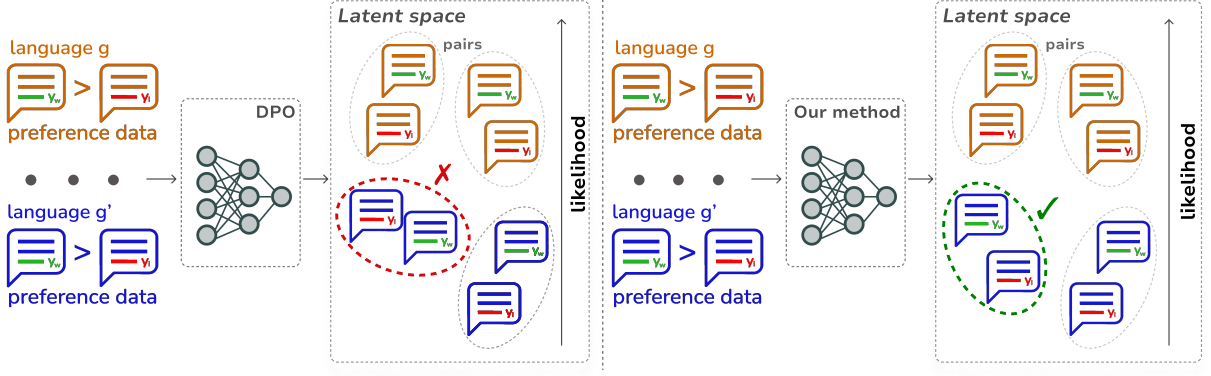


Figure 1: Comparison of our method with DPO. On the left side of the figure, DPO shows some pairs that are misaligned or have ambiguous alignment (✗). Our method fixes this, as shown on the right side (✓), where the model aligns well in both languages. Best viewed in color. Preference data icons from (Rafailov et al., 2023).

Their results show that DPO reduces toxicity when fine-tuned with this goal, but toxicity levels remain unequal across languages.

This work aims to reduce cross-lingual disparities in aligning LLMs with human values. To this end, we regularize DPO training with the difference in model perplexities across languages. Experimental results show a reduction of the alignment gap with our strategy.

The rest of the paper is organized as follows. Section 2 provides the details on our method. Sections 3 and 4 respectively introduce the setting of our experiments and the results. Finally, the last section concludes the paper and opens the way for future research.

2 Proposition

Our objective is to align LLMs with human values using DPO, while simultaneously reducing the alignment gap across languages. We provide the background knowledge related with DPO in Appendix A. It is important to clarify that we do not aim to enforce similar alignment across languages. Indeed, norms and values can differ across cultures (Ramezani and Xu, 2023). Thus, our objective is to induce similar levels of alignment across languages without erasing inherent cultural variations. To this end, we modify the DPO loss by adding a regularization term that penalizes differences in perplexity between continuations in different languages. We illustrate the problem and our proposition in Figure 1. Let $\mathcal{D} = \{x_i, y_{w,i}, y_{l,i}, g_i\}_i^N$ be a dataset of size N with x_i a sequence, $y_{w,i}$ the preferred continuation, $y_{l,i}$ the rejected continuation, and g_i the

language of the instance. Let $\pi_\theta(y | x)$ be the log-likelihood of the continuation y given the sequence x for the trained model. We note:

$$\rho_\theta(y | x) = \exp\left(\frac{-\pi_\theta(y | x)}{n_y}\right),$$

the perplexity of the model for y given x with n_y the number of tokens in the continuation. Then, our regularization term is:

$$\mathcal{L}_{\text{REG}} = \mathbb{E}_{(x, y_w, y_l, g) \sim \mathcal{D}} \left[\frac{1}{n_j} \sum_{j: g_j \neq g_i} \left| \frac{\rho_\theta(y_{w,i} | x_i)}{\rho_\theta(y_{l,i} | x_i)} - \frac{\rho_\theta(y_{w,j} | x_j)}{\rho_\theta(y_{l,j} | x_j)} \right| \right],$$

with $n_j = |\{j | g_j \neq g_i\}|$.

The motivation for computing the gap between the perplexity ratios of preferred and rejected continuations is to enforce cross-lingual preference consistency. The DPO objective increases the likelihood of preferred continuations while decreasing that of rejected ones. However, as shown in Appendix B, likelihoods and perplexities for semantically equivalent sentences may vary across languages. Consequently, the ratio $\frac{\rho_\theta(y_w|x)}{\rho_\theta(y_l|x)}$ can be interpreted as a measure of preference strength, reflecting how much more likely the model considers the preferred continuation relative to the rejected one. If this preference strength differs across languages for equivalent examples, it indicates an inconsistency in the model’s cross-lingual behavior. Thus, we penalize the DPO loss by this difference to encourage consistent relative preference patterns across languages, independent of language-specific

likelihood scaling, while preserving alignment performance within each language. Importantly, this approach does not enforce identical preference values across languages or cultures, but rather promotes comparable levels of alignment across them.

Moreover, we consider perplexities rather than the log-likelihoods of the sequences, as in DPO, due to the impact of the sequence length on these measures, as outlined by Dong et al. (2024), who show that longer sequences receive higher likelihoods and rewards. Note that our regularization term computes a ratio of per-sequence perplexities, normalizing each sequence’s length independently rather than jointly length-normalizing the pairs. Empirically, computing perplexity tends to mitigate the impact of length on the reward computation. However, in extreme cases where the two sequences differ substantially in length, a unified normalization would likely be preferable.

Finally, our training loss is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{DPO}} + \gamma \cdot \mathcal{L}_{\text{REG}},$$

with γ a hyperparameter controlling the strength of the regularization term.

In non-parallel settings, that is, when corresponding entries from different languages are unavailable, our regularization term operates at the distributional level rather than enforcing per-example cross-lingual consistency: for a given instance, the perplexity ratio is compared against instances from other languages without semantic pairing but with similar preferences. This way, we reduce the global alignment gap between languages, ensuring the model does not exhibit systematically different alignment behavior depending on the language, rather than aligning individual semantically equivalent examples. When parallel corpora are available, semantically equivalent pairs across languages naturally provide this per-example correspondence.

In the next section, we describe the evaluation protocol used to assess the impact of our approach on LLMs’ alignment.

3 Experiments

3.1 Protocol

We conduct experiments on two LLMs: Mistral-7B-v0.3 (Jiang et al., 2023) and LLaMA3-8B (Grattafiori et al., 2024), respectively referred to as Mistral and LLaMA models in the following.

We compare our method (dpo_{ppi}) against standard DPO fine-tuning (dpo) and the original pre-

trained models without fine-tuning (base). We perform 3-fold cross-validation, described in Appendix E, to select optimal hyperparameters for dpo and dpo_{ppi}. We report the computational information in Appendix F.

3.2 Datasets

We use two corpora that *may contain offensive content* to assess the alignment of LLMs with human values. We study toxicity and the moral alignment of LLMs. Regarding languages, we compare English with French, Russian, and Spanish. As the LLMs we study are mostly trained in English, we consider it as the reference language.

The Multilingual Jigsaw dataset (Kivlichan et al., 2020) includes online comments labeled for toxicity. We label toxic content as *rejected* and non-toxic content as *preferred*. Preference pairs are formed randomly. For each language, the corpus contains 2000 training and 1340 test pairs.

The Moral corpus combines the MoralStories (Emelin et al., 2021) and HistoiresMorales (Leteno et al., 2025) datasets for English and French. These parallel datasets include short stories (x) followed by a moral action (y_w) and an immoral action (y_l). MoralStories and HistoiresMorales each contain 8400 training and 3600 test pairs. We also consider the UniMoral dataset (Kumar and Jurgens, 2025) to compare English with Russian and Spanish, but the dataset size (488 examples per language) is not sufficient to align the models properly. We present this analysis in Appendix C. During training, we use half of each dataset’s training set without overlap, ensuring that each story appears only once.

3.3 Metrics

We evaluate the effectiveness of our approach in reducing the alignment gap across languages during training by measuring model perplexity on sequence pairs. Following Nangia et al. (2020); Névéol et al. (2022) for fairness and Leteno et al. (2025) for morality, perplexity serves as an indicator of model uncertainty where lower values indicate better alignment. From this, we define the Preference Rate (PR) over a test set of size N_T as:

$$PR = \frac{|\rho_\theta(y_w|x) \leq \rho_\theta(y_l|x)|}{N_T}.$$

We denote PR_G and $PR_{\bar{G}}$ as the Preference Rates for languages G and $\bar{G}$, respectively. We then define the alignment gap as $\Delta = |PR_G - PR_{\bar{G}}|$.

Language	Loss	γ	β	$PR_{EN} (\uparrow)$	$PR_X (\uparrow)$	$\Delta (\downarrow)$	$S_{PR}(\uparrow)$
Mistral							
en-fr	base	-	-	59.81 ± 0.0	64.59 ± 0.0	4.78 ± 0.0	0.0
	dpo	0	0.001	93.83 ± 0.86	85.96 ± 1.54	7.87 ± 0.86	0.95
	dpo _{ppl}	0.01	0.001	94.36 ± 0.75	86.8 ± 1.32	7.56 ± 0.92	1.1
en-es	base	-	-	59.81 ± 0.0	66.51 ± 0.0	6.7 ± 0.0	1.0
	dpo	0	0.1	92.85 ± 0.80	87.03 ± 0.75	5.81 ± 1.12	1.23
	dpo _{ppl}	1.0	0.001	92.97 ± 3.32	88.5 ± 1.66	4.47 ± 1.95	1.69
en-ru	base	-	-	59.81 ± 0.0	70.1 ± 0.0	10.29 ± 0.0	1.0
	dpo	0	0.01	94.14 ± 0.47	87.89 ± 1.16	6.25 ± 0.78	1.55
	dpo _{ppl}	0.1	0.01	92.84 ± 0.39	89.12 ± 0.48	3.73 ± 0.47	1.96
LLaMA							
en-fr	base	-	-	63.64 ± 0.0	64.95 ± 0.0	1.31 ± 0.0	0.0
	dpo	0	0.001	94.86 ± 0.41	86.27 ± 0.94	8.59 ± 0.92	1.0
	dpo _{ppl}	0.1	0.01	92.94 ± 0.31	85.36 ± 0.79	7.58 ± 0.81	1.03
en-es	base	-	-	63.64 ± 0.0	64.47 ± 0.0	0.83 ± 0.0	0.0
	dpo	0	0.01	93.2 ± 0.56	87.85 ± 0.61	5.36 ± 0.5	0.98
	dpo _{ppl}	0.01	0.01	93.4 ± 0.32	88.26 ± 0.46	5.14 ± 0.41	1.05
en-ru	base	-	-	63.64 ± 0.0	67.11 ± 0.0	3.47 ± 0.0	0.09
	dpo	0	0.01	93.32 ± 0.71	89.14 ± 1.03	4.19 ± 0.99	0.97
	dpo _{ppl}	0.001	0.01	93.28 ± 0.41	89.88 ± 0.71	3.4 ± 0.58	2.0

Table 1: Results on the Jigsaw dataset, averaged over 5 seeds. Best results are shown in bold.

Language	Loss	γ	β	$PR_{EN} (\uparrow)$	$PR_X (\uparrow)$	$\Delta (\downarrow)$	$S_{PR}(\uparrow)$
Mistral							
en-fr	base	-	-	43.86 ± 0.0	47.92 ± 0.0	4.06 ± 0.0	1.0
	dpo	0	0.01	98.99 ± 0.1	97.21 ± 0.35	1.78 ± 0.33	1.82
	dpo _{ppl}	0.1	0.25	99.02 ± 0.22	97.69 ± 0.2	1.33 ± 0.09	2.0
LLaMA							
en-fr	base	-	-	46.72 ± 0.0	48.75 ± 0.0	2.03 ± 0.0	1.0
	dpo	0	0.01	98.88 ± 0.16	97.82 ± 0.22	1.07 ± 0.14	1.96
	dpo _{ppl}	1.0	0.1	98.71 ± 0.16	97.68 ± 0.22	1.03 ± 0.23	1.99

Table 2: Results on the Moral corpus, averaged over 5 seeds. Best results are shown in bold.

Lastly, to evaluate the trade-off between performance on each language and gap minimization, we use an aggregate score combining these factors:

$$S_{PR} = n_{[0,1]}(PR_G) + n_{[0,1]}(PR_{\bar{G}}) - n_{[0,1]}(\Delta),$$

with $n_{[0,1]}(\cdot)$ is the min-max normalization.

4 Results

We assess LLMs’ alignment with human values by analyzing sequence perplexities. Results for the Jigsaw dataset and for the Moral corpus are respectively presented in Tables 1 and 2. First, both dpo and dpo_{ppl} significantly improve LLMs’ alignment compared to the base models, achieving comparable levels of alignment. Importantly, our method

(dpo_{ppl}) consistently shows the highest trade-off score S_{PR} across all datasets and models. This indicates that our approach effectively balances alignment and fair treatment for both languages. Moreover, the alignment gap Δ is systematically reduced with dpo_{ppl} compared to dpo. We observe that this reduction is larger on Jigsaw than on the Moral corpus. We attribute this difference to the data structure: the Moral corpus is parallel, with the French, HistoiresMorales, being a translation of the English, MoralStories, leading to highly similar distributions and thus a smaller initial gap. In contrast, Jigsaw contains online comments from more diverse domains, increasing cross-lingual discrepancies. Thus, our method yields a stronger alignment

gap reduction in this more heterogeneous setting. Moreover, since Jigsaw is non-parallel, topic, style, and base-rate differences across languages are not fully disentangled from language effects; the larger reduction observed on Jigsaw should therefore be interpreted alongside this broader source of cross-lingual variation.

Interestingly, we observe that the regularization can also improve alignment on English data itself, despite English not being the primary target of the cross-lingual objective. This finding is consistent with a growing body of work in multilingual NLP showing that training across languages acts as a form of cross-lingual regularization: by encouraging the model to behave consistently across languages, it is pushed toward more robust, language-agnostic representations, which in turn can also benefit performance on the source language (Gogoulou et al., 2022).

Finally, we control the model’s general abilities post-alignment in Appendix D.

5 Conclusion

In this paper, we propose a novel method to align LLMs with human values while reducing the alignment gap between languages. We introduce a regularization term during DPO training that penalizes discrepancies in model scores across languages. We demonstrate that this regularized fine-tuning moves models closer to the goal of equitable alignment across languages, consistently reducing alignment gaps on all models and datasets evaluated.

Despite these promising results, our approach still has room for improvement. In particular, in the non-parallel case, the distributional objective discussed in Section 2 cannot fully disentangle language effects from domain or content differences across examples. Therefore, enforcing per-example cross-lingual consistency, for instance by identifying or constructing close cross-lingual pairs, would constitute an interesting direction for future work.

More broadly, unequal alignment across languages can result in prejudice against users based on their language, a concern that persists despite the widespread availability of LLM-based tools. Our method does not fully close the alignment gap but represents a step toward addressing it.

Future work should also examine models’ allocational performance in decision-making scenarios and whether disparities can also be effectively reduced in such applications.

Limitations

While we investigate the alignment of LLMs with human values, we recognize that this approach is inherently constrained by the nature of our datasets. These datasets do not comprehensively capture the full spectrum of human values, which are diverse, context-dependent, and culturally variable. Consequently, our conclusions about value alignment should be interpreted with caution and seen as preliminary steps toward a more inclusive and representative assessment.

Ethical considerations

In this paper, we study the alignment of LLMs with human values, leveraging existing benchmark datasets. We acknowledge that some of these datasets may include harmful or sensitive content. However, such data is used within a scientific context to facilitate meaningful analysis. We propose a method aimed at improving model alignment with values broadly recognized as moral, non-toxic, and socially acceptable. It is important to emphasize that this method is not intended and should not be used for purposes opposite to this objective, such as reinforcing harmful or unethical behaviors. While there is an inherent risk of misuse, we believe our research contributes positively to the development of safer, more responsible language technologies that better serve users across different linguistic and cultural backgrounds.

Acknowledgments

This work was funded by the French National Agency for Research (ANR) in the context of the Diké project (ANR-21-CE23-0026) and is supported by Hi! PARIS and ANR/France 2030 program (ANR-23-IACL-0005). We also benefited from the computational resources of the Hubert Curien Laboratory (Jean Monnet University, Saint-Etienne) for this work.

References

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. [The multilingual alignment prism: Aligning global and local preferences to reduce harm](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12027–12049, Miami, Florida, USA. Association for Computational Linguistics.

- Utkarsh Agarwal, Kumar Tanmay, Aditi Khandelwal, and Monojit Choudhury. 2024. [Ethical reasoning and moral value alignment of LLMs depend on the language we prompt them in](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6330–6340, Torino, Italia. ELRA and ICCL.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Alexander Bick, Adam Blandin, and David J Deming. 2026. The rapid adoption of generative ai. *Management Science*.
- Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. [Systematic inequalities in language technology performance across the world’s languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.
- 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.
- Chad Coleman, W Russell Neuman, Ali Dasdan, Saifinah Ali, and Manan Shah. 2025. The convergent ethics of ai? analyzing moral foundation priorities in large language models with a multi-framework approach. *arXiv preprint arXiv:2504.19255*.
- Benoit Courty, Victor Schmidt, Sasha Luccioni, Goyal-Kamal, MarionCoutarel, Boris Feld, Jérémy Lecourt, LiamConnell, Amine Saboni, Inimaz, supatomic, Mathilde Léval, Luis Blanche, Alexis Cruveiller, ouminasara, Franklin Zhao, Aditya Joshi, Alexis Bogroff, Hugues de Lavoreille, and 11 others. 2024. [mlco2/codecarbon: v2.4.1](#).
- Michael Han Daniel Han and Unsloth team. 2023. [Unsloth](#).
- Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf. *Transactions on Machine Learning Research*.
- Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. [Moral stories: Situated reasoning about norms, intents, actions, and their consequences](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Evangelia Gogoulou, Ariel Ekgren, Tim Isbister, and Magnus Sahlgren. 2022. [Cross-lingual transfer of monolingual models](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 948–955, Marseille, France. European Language Resources Association.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Katharina Hämmerl, Bjoern Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin Rothkopf, Alexander Fraser, and Kristian Kersting. 2023. [Speaking multiple languages affects the moral bias of language models](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2137–2156, Toronto, Canada. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. [Aligning ai with shared human values](#). In *International Conference on Learning Representations*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Jiwoo Hong, Noah Lee, Rodrigo Martínez-Castaño, César Rodríguez, and James Thorne. 2024. Cross-lingual transfer of reward models in multilingual alignment. *arXiv preprint arXiv:2410.18027*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo 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.
- Ian Kivlichan, Jeffrey Sorensen, Julia Elliott, Lucy Vasserman, Martin Görner, and Phil Culliton. 2020. Jigsaw multilingual toxic comment classification. <https://kaggle.com/competitions/jigsaw-multilingual-toxic-comment-classification>. Kaggle.
- Shivani Kumar and David Jurgens. 2025. [Are rules meant to be broken? understanding multilingual moral reasoning as a computational pipeline with UniMoral](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5890–5912, Vienna, Austria. Association for Computational Linguistics.
- Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. [ChatGPT beyond English: Towards a comprehensive evaluation of large](#)

- language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.
- Thibaud Leteno, Irina Proskurina, Antoine Gourru, Julien Velcin, Charlotte Laclau, Guillaume Metzler, and Christophe Gravier. 2025. **HISTOIRES-MORALES: A French dataset for assessing moral alignment**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2590–2612, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xiaochen Li, Zheng Xin Yong, and Stephen Bach. 2024. Preference tuning for toxicity mitigation generalizes across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13422–13440.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. **CrowS-pairs: A challenge dataset for measuring social biases in masked language models**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.
- Aur lie N v  ol, Yoann Dupont, Julien Bezan  on, and Kar  n Fort. 2022. **French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.
- 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.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 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. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.
- Aida Ramezani and Yang Xu. 2023. **Knowledge of cultural moral norms in large language models**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. Association for Computational Linguistics.
- Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. **MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand. Association for Computational Linguistics.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, and 1 others. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.
- Alejandro Tlaie. 2024. **Exploring and steering the moral compass of large language models**. *Preprint*, arXiv:2405.17345.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth  e Lacroix, Baptiste Rozi  re, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- 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 *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1332–1353.
- Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, and Deyi Xiong. 2024. **Exploring multilingual concepts of human values in large language models: Is value alignment consistent, transferable and controllable across languages?** In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1771–1793, Miami, Florida, USA. Association for Computational Linguistics.
- Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025. Implicit cross-lingual rewarding for efficient multilingual preference alignment. *arXiv preprint arXiv:2503.04647*.

A Direct Preference Optimization

DPO directly optimizes model parameters from preference data, rather than relying on explicitly defined reward functions. This suppresses the need to train a separate reward model during the fine-tuning. DPO assumes access to a set of pairwise preferences, typically obtained through human feedback. The objective is to align the model’s behavior with these preferences by increasing the likelihood of preferred outputs while decreasing that of rejected ones. The optimization is framed

as a preference-based objective, promoting preferred behaviors over less preferred ones. This is implemented by minimizing a loss function derived from the Bradley-Terry model (Bradley and Terry, 1952), using a softmax-like formulation. Let $\mathcal{D} = \{x_i, y_{w,i}, y_{l,i}\}_i^N$ be a dataset of size N with x_i a sequence, $y_{w,i}$ the preferred continuation, $y_{l,i}$ the rejected continuation. We note σ the sigmoid function, π_θ and π_{ref} the trainable and reference models policy. β is a hyperparameter controlling the deviation of π_θ from the reference model. The reward for a continuation y given an input x is

$$\hat{r}_\theta(x, y) = \beta \cdot \log \frac{\pi_\theta(y | x)}{\pi_{\text{ref}}(y | x)}.$$

Then, the DPO loss is:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = \mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} [\log \sigma(\hat{r}_\theta(x, y_w) - \hat{r}_\theta(x, y_l))].$$

B Variations of likelihood across languages

In this section, we show that for similar sequences in different languages, a model outputs different likelihoods. We consider an example from the Moral corpus in both languages. The prompts and continuations, presented in Table 3, have equivalent length. In Table 4, we observe that for a given model and a similar sentence in French and English, the likelihood of the sentence is different. For all models, we note that it is lower in French than in English.

C Unimoral

To compare the cross-lingual moral alignment of LLMs in English, Spanish, and Russian, we study the UniMoral dataset (Kumar and Jurgens, 2025). The dataset is composed of 488 entries per language, which we split between 188 training examples and 300 test examples. We privileged a larger test set to have enough data to estimate the alignment. However, as noted in Table 5, the training set is insufficient to align LLaMA or Mistral models with human values, compared to the base models. Similarly, the gap reduction is inconsistent with limited training data, despite our method, dpo_{ppl} , having the best trade-off score, S_{PR} , in 3 out of 4 experiments.

Prompts	
English	Mary is working on a group project with other students, she doesn't like some of the work one of the other students has done. Mary wants to turn in the group project and get an A.
French	Marie travaille sur un projet de groupe avec d'autres étudiants, mais elle n'aime pas le travail réalisé par l'un d'entre eux. Marie souhaite obtenir un A pour le projet de groupe.
Continuations	
English	Mary talks with the student about their work and makes some suggestions to improve it.
French	Elle discute avec l'étudiant concerné de son travail et lui fait quelques suggestions pour l'améliorer.

Table 3: Example used for likelihood gap evaluation across languages.

Model	English		French	
	$\pi_\theta(y x)$	$\rho_\theta(y x)$	$\pi_\theta(y x)$	$\rho_\theta(y x)$
Mistral	3.8	43.7	2.5	11.9
LLaMA	4.1	60.1	2.6	13.13

Table 4: Results of the likelihood $\pi_\theta(y | x)$ and perplexity $\rho_\theta(y | x)$ comparison for different languages on Mistral-7B-v0.3, LLaMA-3.1-8B.

D Impact on General Natural Language Understanding

In this section, we investigate the impact of the alignment procedures on the general language understanding abilities of the models. To do so, we consider the models' performance on the Massive Multitask Language Understanding benchmark (Hendrycks et al., 2020). The benchmark measures the model's abilities on various tasks, such as maths, question answering on history, or philosophy. We present results in Table 7. We note close results, within 1%, between the base model and the model after dpo or dpo_{ppl} training. This highlights that our approach and dpo improve the models' alignment without negative impact on the general abilities.

Language	Loss	γ	β	$PR_{EN} (\uparrow)$	$PR_X (\uparrow)$	$\Delta (\downarrow)$	$S_{PR}(\uparrow)$
Mistral							
en-es	base	-	-	50.0 $\pm$ 0.0	46.33 $\pm$ 0.0	3.67 $\pm$ 0.0	0.0
	dpo	0	0.001	49.6 $\pm$ 0.49	50.87 $\pm$ 0.78	1.27 $\pm$ 0.65	0.86
	dpo _{ppl}	0.001	1.0	49.93 $\pm$ 0.49	49.2 $\pm$ 0.75	0.87 $\pm$ 0.81	1.46
en-ru	base	-	-	50.0 $\pm$ 0.0	51.33 $\pm$ 0.0	1.33 $\pm$ 0.0	1.0
	dpo	0	0.25	49.73 $\pm$ 0.61	53.6 $\pm$ 0.88	3.87 $\pm$ 1.31	0.54
	dpo _{ppl}	0.01	0.1	49.47 $\pm$ 0.34	54.13 $\pm$ 1.05	4.67 $\pm$ 1.19	0.0
LLaMA							
en-es	base	-	-	49.33 $\pm$ 0.0	46.67 $\pm$ 0.0	2.66 $\pm$ 0.0	1.0
	dpo	0	0.5	49.47 $\pm$ 0.27	47.67 $\pm$ 0.21	1.8 $\pm$ 0.34	0.93
	dpo _{ppl}	1.0	0.001	50.33 $\pm$ 0.47	47.93 $\pm$ 0.49	2.4 $\pm$ 0.53	1.3
en-ru	base	-	-	49.33 $\pm$ 0.0	49.67 $\pm$ 0.0	0.34 $\pm$ 0.0	0.0
	dpo	0	0.001	49.47 $\pm$ 0.17	50.13 $\pm$ 0.45	0.8 $\pm$ 0.34	0.52
	dpo _{ppl}	0.001	0.5	49.6 $\pm$ 0.14	49.87 $\pm$ 0.34	0.4 $\pm$ 0.25	1.3

Table 5: Results on the Moral corpus, averaged over 5 seeds. The best results between dpo and dpo_{ppl} are shown in bold. base is given as a baseline in terms of PR .

E Experimental details

To fine-tune the models, we adapt HuggingFace DPOTrainers² and use the Unsloth library (Daniel Han and team, 2023) for QLoRA training.

Hyperparameters	dpo	dpo _{ppl}
γ	0	1e ⁻³ , 1e ⁻² , 1e ⁻¹ , 1.0
β		1e ⁻³ , 1e ⁻² , 1e ⁻¹ , 1.0

Table 6: Hyperparameters tested in the cross-validation for the representational evaluation.

We realize a 3-fold cross-validation testing the hyperparameters shown in Table 6, and present the best ones in the result tables. For the models, we use the versions available on HuggingFace:

- Mistral-7b: <https://huggingface.co/mistralai/Mistral-7B-v0.3>
- LLaMA3-8b: <https://huggingface.co/meta-LLaMA/LLaMA-3.1-8B>

We select the best hyperparameters for each method based on the trade-off score S_{PR} .

F Computational cost

We conduct the experiments on Nvidia A40 and L4 GPUs, respectively, with 48Go and 24Go of RAM. Moreover, using the CodeCarbon library (Courty et al., 2024), we report the carbon emissions, electricity cost, and running time for one training in

²https://huggingface.co/docs/trl/main/en/dpo_trainer#trl.DPOTrainer

Table 8. We report the information on the runs for the English-French language pairs.

G License

Dataset The MoralStories and HistoiresMorales datasets are available under the MIT license. The Jigsaw dataset is available under the Creative Commons license. No license is specified for the Unimoral dataset.

Model The Mistral is available under the Apache 2.0 license, and the LLaMA model is available under the LLaMA 3.1 Community License. The Unsloth library and HuggingFace trainers are available under the Apache 2.0 license.

Writing LLMs were used as a writing assistant tool for fluency purposes, translation, and typo correction.

Datasets & languages	Loss	MMLU	humanities	other	social sciences	stem
Mistral						
-	base	0.5963 ± 0.0039	0.5330 ± 0.0067	0.6846 ± 0.0081	0.6968 ± 0.0081	0.5056 ± 0.0086
Moral						
en-fr	dpo	0.5959 ± 0.0039	0.5454 ± 0.0068	0.6707 ± 0.0081	0.6929 ± 0.0081	0.5030 ± 0.0086
	dpo _{ppl}	0.6008 ± 0.0039	0.5518 ± 0.0068	0.6827 ± 0.0081	0.6974 ± 0.0080	0.4992 ± 0.0086
Jigsaw						
en-fr	dpo	0.5919 ± 0.0039	0.5294 ± 0.0067	0.6752 ± 0.0081	0.6939 ± 0.0081	0.5036 ± 0.0086
	dpo _{ppl}	0.5915 ± 0.0039	0.5322 ± 0.0068	0.6778 ± 0.0081	0.6877 ± 0.0081	0.5011 ± 0.0086
en-es	dpo	0.5925 ± 0.0039	0.5322 ± 0.0067	0.6785 ± 0.0081	0.6870 ± 0.0081	0.5056 ± 0.0086
	dpo _{ppl}	0.5902 ± 0.0039	0.5333 ± 0.0067	0.6691 ± 0.0082	0.6890 ± 0.0081	0.5008 ± 0.0086
en-ru	dpo	0.5956 ± 0.0039	0.5386 ± 0.0068	0.6749 ± 0.0081	0.6909 ± 0.0081	0.5094 ± 0.0086
	dpo _{ppl}	0.5955 ± 0.0039	0.5382 ± 0.0068	0.6772 ± 0.0081	0.6909 ± 0.0081	0.5075 ± 0.0086
LLaMA						
-	base	0.6404 ± 0.0038	0.5792 ± 0.0067	0.7119 ± 0.0078	0.7481 ± 0.0077	0.5563 ± 0.0085
Moral						
en-fr	dpo	0.6345 ± 0.0038	0.5681 ± 0.0066	0.7155 ± 0.0078	0.7403 ± 0.0077	0.5506 ± 0.0085
	dpo _{ppl}	0.6339 ± 0.0038	0.5683 ± 0.0067	0.7142 ± 0.0078	0.7377 ± 0.0077	0.5512 ± 0.0085
Jigsaw						
en-fr	dpo	0.6377 ± 0.0038	0.5722 ± 0.0066	0.7116 ± 0.0078	0.7468 ± 0.0077	0.5563 ± 0.0085
	dpo _{ppl}	0.6371 ± 0.0038	0.5739 ± 0.0067	0.7107 ± 0.0078	0.7420 ± 0.0077	0.5566 ± 0.0085
en-es	dpo	0.6400 ± 0.0038	0.5798 ± 0.0067	0.7152 ± 0.0078	0.7433 ± 0.0077	0.5550 ± 0.0085
	dpo _{ppl}	0.6330 ± 0.0038	0.5679 ± 0.0066	0.7084 ± 0.0079	0.7423 ± 0.0077	0.5490 ± 0.0085
en-ru	dpo	0.6415 ± 0.0038	0.5862 ± 0.0067	0.7132 ± 0.0078	0.7436 ± 0.0077	0.5538 ± 0.0086
	dpo _{ppl}	0.6411 ± 0.0038	0.5843 ± 0.0067	0.7113 ± 0.0078	0.7449 ± 0.0077	0.5557 ± 0.0085

Table 7: Results on the MMLU benchmark.

Model	Dataset	Loss	Emissions (CO ₂ eq)	Electricity Consumption (kWh)	Duration (seconds)	Device	GPU RAM
Mistral-7B-v0.3	Moral	DPO	0.060	0.3567	1747	NVIDIA A40	48Go
Mistral-7B-v0.3	Moral	DPO _{ppl}	0.085	0.5006	3035	NVIDIA L4	24Go
Mistral-7B-v0.3	Jigsaw	DPO	0.017	0.1001	541	NVIDIA L4	24Go
Mistral-7B-v0.3	Jigsaw	DPO _{ppl}	0.034	0.1996	990	NVIDIA A40	48Go
Llama-3.1-8B	Moral	DPO	0.059	0.3490	1825	NVIDIA A40	48Go
Llama-3.1-8B	Moral	DPO _{ppl}	0.059	0.3473	1815	NVIDIA A40	48Go
Llama-3.1-8B	Jigsaw	DPO	0.033	0.1973	1091	NVIDIA A40	48Go
Llama-3.1-8B	Jigsaw	DPO _{ppl}	0.034	0.2005	1098	NVIDIA A40	48Go

Table 8: Training pipeline computational costs.