RLHF Can Speak Many Languages: Unlocking Multilingual Preference Optimization for LLMs

John Dang[1], Arash Ahmadian[1,2], Kelly Marchisio[2] Julia Kreutzer[1], Ahmet Üstün[1], Sara Hooker[1]

[1]Cohere For AI, [2]Cohere
{johndang,arash,kelly,juliakreutzer,ahmet,sarahooker}@cohere.com

Abstract

Preference optimization techniques have become a standard final stage for training state-of-art large language models (LLMs). However, despite widespread adoption, the vast majority of work to-date has focused on a small set of high-resource languages like English and Chinese. This captures a small fraction of the languages in the world, but also makes it unclear which aspects of current state-of-the-art research transfer to a multilingual setting.

In this work, we perform an exhaustive study to achieve a new state of the art in aligning multilingual LLMs. We introduce a novel, scalable method for generating high-quality multilingual feedback data to balance data coverage. We establish the benefits of cross-lingual transfer and increased dataset size in preference training. Our preference-trained model achieves a 54.4% win-rate against Aya 23 8B, the current state-of-the-art multilingual LLM in its parameter class, and a 69.5% win-rate or higher against widely used models like Gemma, Mistral and Llama 3. As a result of our efforts, we expand the frontier of alignment techniques to 23 languages, covering approximately half of the world's population.

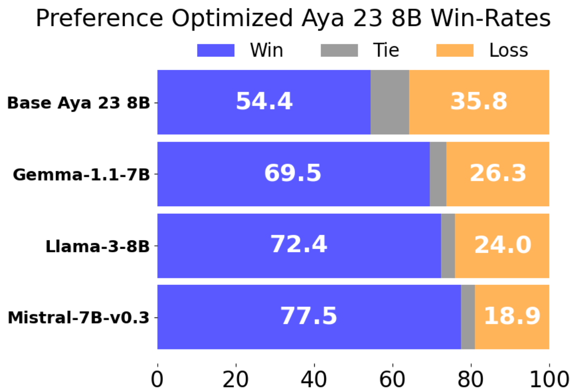


Figure 1: **Win-rates between our preference-trained model with the other state-of-the-art open weight models**: Our preference-trained model based on Aya-23-8B significantly outperforms the original Aya-23-8B, Gemma-1.1-7B-it, Meta-Llama3-8B-Instruct, and Mistral-7B-Instruct-v0.3. Win-rates are computed with GPT-4-Turbo as a judge.

1 Introduction

What languages are favored in technological progress is often deeply intertwined with historical patterns of technology access and resources (∀ et al., 2020; Bird, 2022; Singh et al., 2024). Preference optimization is a valuable and widely adopted post-training technique to align large language models (LLMs) with human preferences (Christiano et al., 2017b; Stiennon et al., 2022; Ouyang et al., 2022a; Bai et al., 2022). It has also been shown to lead to large gains in performance across a wide variety of NLP tasks (Wang et al., 2024; Ivison et al., 2023; Xu et al., 2024; Lightman et al., 2024). To-date, the majority of progress in preference optimization has over-fit to a small handful of languages, resulting in large gaps in performance outside of English (Schwartz et al., 2022; Kotek et al., 2023; Khandelwal et al., 2023; Vashishtha et al., 2023; Khondaker et al., 2023), and also risks introducing security flaws for all users (Yong et al., 2024; Nasr et al., 2023; Li et al., 2023a; Lukas et al., 2023; Deng et al., 2023).

While expanding the frontier of what languages are supported by AI is an increasingly urgent problem, extending preference optimization to a multilingual setting is a non-trivial challenge. First, numerous works have shown that multilingual modeling typically faces both a data scarcity and data quality problem (Singh et al., 2024; Üstün et al., 2024; Dodge et al., 2021; Kreutzer et al., 2022; Luccioni and Viviano, 2021). This is even more pronounced for high-quality preference data which is virtually non-existent in many languages. Collecting multilingual preference datasets through human annotation is expensive and time intensive (Boubdir et al., 2023; Chaudhari et al., 2024), and while

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 13134–13156
November 12-16, 2024 ©2024 Association for Computational Linguistics

prior works have proposed the use of LLMs to synthetically create preference datasets (Bai et al., 2022; Yuan et al., 2024; Pace et al., 2024), these efforts predominantly focus on English. The few efforts that have focused on multilinguality have relied on translation, introducing artifacts and resulting in a lack of diversity in preference pairs (Lai et al., 2023), which is known to be critical to model performance (Naik et al., 2023; Chung et al., 2023; Li et al., 2023b; Lahoti et al., 2023; Kirk et al., 2024a).

Second, preference optimization in many languages simultaneously is a difficult machine learning task. The lack of studies in preference optimization outside of English raises questions on how findings from monolingual optimization would transfer. Training dynamics of RLHF are known to be often unstable (Casper et al., 2023; Gao et al., 2022a; Chaudhari et al., 2024), which can be exacerbated by the involvement of multiple languages where preference data is from a heterogeneous distribution and negative transfer between languages is possible (Wang et al., 2020, 2019). The few existing works on multilingual RLHF (Lai et al., 2023, 2024) exhibit poor results and are outperformed by massively multilingual language models without any preference optimization (Üstün et al., 2024; Aryabumi et al., 2024).

Poor performance of the few existing multilingual preference optimization works begs the question: *Is this a result of fundamental limitations with standard preference optimization techniques (especially in heterogeneous optimization settings like multilingual) or whether we are lacking high quality multilingual data?*

In this work, we exhaustively study the aforementioned challenges. Our goal is to systematically understand key variables which might impact multilingual alignment, including the source and amount of available preference data, offline vs online RLHF techniques, and the effect varying number of languages covered in preference optimization training data. We complete a comprehensive set of experiments with state-of-the-art alignment techniques including DPO (Rafailov et al., 2023) and RLOO (Kool et al., 2019; Ahmadian et al., 2024) starting from the 8-billion-parameter instruction-finetuned Aya model covering 23 languages (Aya-23-8B; Aryabumi et al., 2024). Our findings can be summarized as follows:

1. **Preference optimization exhibits cross-lingual transfer**. We show that preference-optimization even with only English data improves performance in other languages. However, the addition of a few more languages significantly increases cross-lingual transfer, achieving win-rates on unseen languages up to 54.9% when including 5 languages in training data compared to 46.3% when training only on English.

2. **Multilingual preference data is necessary for aligning multilingual LLMs**. We find that increasing the number of languages in preference optimization training data consistently improves multilingual performance compared to English-only training data, increasing win-rates by up to 7.0% from 46.4% to 53.4% when all languages are included.

3. **Online preference optimization outperforms offline optimization**. RLOO as an online method achieves better overall performance than DPO by a maximum 10.6% difference in their average win-rates (54.4% vs 43.8%). Furthermore, we find that RLOO also enables better cross-lingual transfer, achieving up to 8.3% increase over DPO in average win-rate on languages not included in training.

4. **Preference optimized Aya 23 8B outperforms other open weights models** Preference optimization leads to large gains in win-rates against both the original Aya model (54.4% win-rate) and widely used models including Meta-Llama3-8B-Instruct (AI@Meta, 2024) (72.4% win-rate), Gemma-1.1-7B-it (Gemma-Team, 2024) (69.5% win-rate), Mistral-7b-Instruct-v0.3 (77.5%win-rate) (Jiang et al., 2023) across all 23 languages as shown in Figure 1.

2 Methodology

2.1 Addressing Data Scarcity

The limited prior work on multilingual preference training involved fully translated preferences from English (Lai et al., 2023), however the Okapi model has since been outperformed by non-preference aligned models including the base Aya 23 8B model we experiment with here (Aryabumi et al., 2024). We hypothesize that the poor performance may be due to the reliance on translated preferences. While language coverage may be improved by translation (Ranaldi and Pucci, 2023; Üstün et al., 2024), the introduction of translation artefacts known as *translationese* (Bizzoni et al., 2020; Vanmassenhove et al., 2021) can hurt performance. Furthermore, repeatedly translating the same preference pairs can hurt preference diversity. The exact trade-off between the positive and negative benefits is not well understood and is difficult to isolate empirically (Yu et al., 2022; Dutta Chowdhury et al., 2022)

Here, we attempt to avoid some of the issues with translation by creating preference pairs that intentionally aim to steer model generations away from *translationese*. First, we construct a diverse set of general instruction-following multilingual prompts by translating approximately 50K English prompts from ShareGPT[1] into the remaining 22 languages supported by Aya 23 8B. Automatic translation is done by using NLLB 3.3B (NLLB Team et al., 2022).

Source of completions After translating prompts, we generate completions for each language by using multiple LLMs of varying multilingual capability. This ensures increased completion diversity versus the alternative of simply translating the original English preference models. More specifically, we use Cohere's Command[2] and Command R+[3] models, where the latter is explicitly trained for multilingual performance.[4] For Command, we generate English completions as the model is primarily proficient in English, and translate them into the other 22 languages. For Command R+, we generate a completion from the same prompt in-language. This method enables obtaining a pair of multilingual completions for each prompt with varying quality based on the difference in models' capabilities and the use of machine translation.

The use of *translated* completions and comparing with high-quality *direct* multilingual generations allows the model to steer away from translation artifacts. Note that the translated completions are ranked as "bad completions" 91% of the time by our reward model annotator, hence, in most cases the preference ranking acts as a proxy label for translated completions.

[1] https://sharegpt.com
[2] https://docs.cohere.com/docs/command-beta
[3] https://docs.cohere.com/docs/command-r-plus
[4] https://docs.cohere.com/docs/command-r-plus#unique-command-r-model-capabilities

2.2 Offline vs Online Preference Training

Reinforcement Learning from Human Feedback (RLHF; Christiano et al., 2017b; Stiennon et al., 2022; Ouyang et al., 2022a; Bai et al., 2022) proposed as the first framework for aligning language models to human preferences, has become a key for training state-of-the-art LLMs (OpenAI et al., 2023; Touvron et al., 2023; Anthropic, 2024; Reid et al., 2024). Canonically, PPO (Schulman et al., 2017b) has been used in RLHF as the online RL algorithm (Stiennon et al., 2022; Ouyang et al., 2022b; Nakano et al., 2021). However, recent offline methods such as Direct Preference Optimization (DPO) (Rafailov et al., 2023) and subsequent works in the same direction (Azar et al., 2024; Ethayarajh et al., 2024; Choi et al., 2024), have proven increasingly popular due to reduced computational complexity. Traditional online methods such as PPO and REINFORCE (Williams, 1992) require an additional network in addition to the policy, maintaining a reward model in memory, and also using the policy to generate doing training, all of which DPO does not require as it is fully offline.

A fractured experimental ecosystem and non-standardized datasets have made it difficult to evaluate the relative merits of both approaches comprehensively. However, recent work in an English setting (Tajwar et al., 2024; Tang et al., 2024) suggests that although the DPO-family of methods theoretically optimize the same objective as traditional RL algorithms, they under-perform compared to well-tuned traditional online RL methods *due to the lack of online generations (on-policy or off-policy) and critique as provided by the reward model*. (Tajwar et al., 2024) also provides empirical evidence that explicit negative gradients during training improve over online methods without them. Given that multilingual datasets are far more heterogeneous than most datasets, it is unclear how these findings apply to massively multilingual settings.

Additionally, it is of great interest what method most benefits from cross-lingual transfer to unseen languages. Thus, in addition to **DPO**, we benchmark **REINFORCE-Leave-one-out** (**RLOO**; Kool et al., 2019; Ahmadian et al., 2024). Ahmadian et al. (2024) shows that PPO may not be the right tool for RLHF, and that simpler REINFORCE-style methods such as Vanilla Policy Gradient and RLOO, are competitive or outperform PPO. In their experiments, RLOO outperforms both PPO and DPO, and incurs significantly lower computational overhead compared to PPO. Additionally, it has a contrastive loss by nature, which (Tajwar et al., 2024) improves learning compared to traditional RL. In Appendix Section A, we give a brief background and introduction to each method.

Preference data mixtures To evaluate if the multilingual preference data is essential in all languages and also to measure the cross-lingual transfer during preference training, we design various preference data mixtures where we control the number of languages covered and the amount of the preference data per language:

1. English-only: English-only preference data mixture that includes 50K prompts with ranked generations. We term this this variant EN-1-50K and use it to understand **cross-lingual benefits that accrue from English-only preferences**. This is important, given it is the standard formulation of research to date.

2. 5 language subset: Multilingual mixture that includes English, Vietnamese, German, Turkish, and Portuguese with a total amount of 50K prompts (10K per language). These 5 languages contain a mixture of higher and lower resource languages, and in their language families and scripts. We refer to this variant as ML-5-50K, training on only this mixture and evaluating on the remaining 18 languages allows us to measure the impact of multilingual preference data on **cross-lingual transfer to unseen languages** in comparison to English-only preference data.

3. All languages (fixed data budget): Multilingual mixture with all 23 languages supported by Aya 23 8B. This is our fixed budget variant where the total number of prompts is kept the same at 50K (approximately 2.2K examples per language) to compare with EN-1-50K and ML-23-50K. This variant **measures performance trade-offs including all languages given the same data budget.** We refer to this variant as ML-23-50K.

4. All languages (not-fixed data budget): Our most comprehensive preference data mixture where we include all 23 languages with 10K prompts per language (230K in total). This mixture which we refer to as ML-23-230K enables us to evaluate the impact of a larger preference data budget in comparison with ML-23-50K.

Agreement between RM and GPT-4	
English	87.9%
Vietnamese	88.7%
Turkish	84.4%
Portuguese	91.0%
German	85.5%
Avg 23 Languages	87.3%

Table 1: Ranking agreement rate for the Reward Model used in our experiments and GPT-4 Turbo on randomly selected Multilingual Dolly responses generated from Command and Command R+ (512 randomly selected per language). Unlike the Reward Model, GPT-4-Turbo is capable of outputting ties. GPT-4-Turbo selects tie result 3.1% of the time on dataset.

3 Experimental Set-up

Multilingual Base Model We perform all experiments with Aya 23 8B (Aryabumi et al., 2024) which is chosen because it **(1)** is massively multilingual, pre-trained and supervised fine-tuned for 23 languages, and **(2)** it achieves state-of-the-art multilingual performance in 23 languages compared to other commonly used LLMs in its class, outperforming Mistral-7B (Jiang et al., 2023), Gemma-7B (Gemma-Team, 2024), and Aya-101-13B (Üstün et al., 2024). On multilingual benchmarks and open-ended generations, Aya 23 achieves a 65% win-rate or higher in head-to-head comparisons with popular open sourcee models(Aryabumi et al., 2024). Furthermore, Aya 23 is an open weights model that is not preference-trained, allowing us to isolate the impact of multilingual preference optimization.

Reward Model We use a closed-source multilingual reward model (RM) which is competitive with top-scoring state-of-the-art RMs on the RewardBench Leaderboard (Lambert et al., 2024). This reward model achieves high LLM response ranking agreement with GPT-4-Turbo on English and Non-English languages as shown in Table 1. We intentionally use a separate reward model from the model we use for llm-as-a-judge evaluation (GPT-4-Turbo[5]) given the known biases incurred by using the same model for both (Verga et al., 2024; Bansal et al., 2024).

Preference Optimization Training We train Aya 23 8B model for 2 epochs in both DPO

[5]https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4

and RLOO experiments. DPO runs are trained with KL-penalty $\beta = 0.5$, learning rate 5e-7, and AdamW optimizer (Kingma and Ba, 2014). RLOO runs are trained with RLOO $k = 2$, KL-penalty $\beta = 0.01$, learning rate $5\text{e}{-6}$, AdamW optimizer, and online generation sampling temperature of 0.75. All runs are performed on a single node with either 8 x Nvidia A100 80GB or 8 x Nvidia H100 80GB GPUs with DeepSpeed ZeRO Stage 3 (Rajbhandari et al., 2020) and full fine-tuning of all 8 billion model parameters. We performed hyperparameter sweeps for learning rate $lr \in \{5\text{e}{-8}, 5\text{e}{-7}, 5\text{e}{-6}, 5\text{e}{-5}\}$ and for KL-penalty $\beta \in \{0.05, 0.1, 0.5\}$ for both DPO and RLOO to the best of our ability. For a fair comparison with DPO, we utilize the same reward model for RLOO which is used to generate our synthetic multilingual preference dataset (for ranking the generations).

3.1 Model Comparisons

We evaluate against multiple state-of-the-art open-source models to ensure a comprehensive evaluation. Details of each model are below:

- **Meta-Llama-3-8B-Instruct** (AI@Meta, 2024) is an 8B parameter open-source instruction fine-tuned model which has been pre-trained on over 15T tokens. Over 5% of the pretraining data is high-quality non-English data, covering over 30 languages. The model is supervised fine-tuned and preference optimized with both offline (DPO) and online (PPO) algorithms
- **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023) is an open-source instruct fine-tuned edition of the Mistral-7B pre-trained model. The model is trained on instruction datasets publicly available on the HuggingFace repository.
- **Gemma-1.1-7B-it** (Gemma-Team, 2024) is a 7B parameter instruction fine-tuned model trained with Gemini models' architectures, data, and training recipes (Gemini-Team et al., 2024) on 6T tokens of data from web documents, mathematics, and code that are primarily English. In addition to the supervised fine-tuning, this model is also further fine-tuned using RLHF on collected pairs of preferences from human annotators.

We note that while the models we evaluate do not explicitly claim to support multiple languages, in practice, they are heavily used by multilingual users relative to explicitly multilingual models like mT0 and BLOOMZ (Muennighoff et al., 2023a). We also observe that they perform well in practice.

3.2 Evaluation

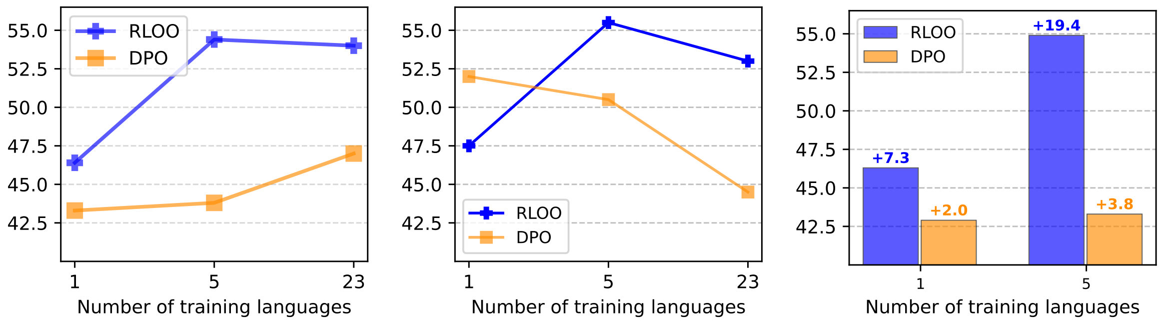
We assess the preference-optimized models on the multilingual open-ended generation and summarization tasks using LLM-simulated evaluation:

1. **Open-ended generations** For open-ended generations, we use dolly-machine-translated test set from the **Aya evaluation suite** (Singh et al., 2024) which is a held-out test set of 200 instances from the Dolly-15k dataset (Conover et al., 2023) translated into 101 languages. This test set was curated by avoiding instructions that include culturally-specific or geographic references. For languages that are available (Arabic, French, Hindi, Russian, and Spanish), we use **dolly-human-edited** test set (Singh et al., 2024), an improved version of **dolly-machine-translated** post-edited by professional human annotators to correct any translation issues.
2. **Summarization Task** For summarization, we use **XLSum** (Hasan et al., 2021), for the subset of 15 languages covered by the benchmark within the Aya 23 language coverage.

Across both tasks, we measure LLM-simulated win-rates which have been shown to be highly correlated with human evaluation for both monolingual English settings (Dubois et al., 2024) and multilingual settings (Üstün et al., 2024). We use GPT-4-Turbo as an LLM-judge and follow the same procedure described by Üstün et al. (2024). To minimize bias, we randomize the order of model outputs. The judge prompt can be found in Appendix D. For evaluation, we use max prompt context length of 512, maximum generation length of 512, and sampling temperature of 0.75.

4 Results and Discussion

Win-rates Against Open-Weights Models Figure 1 and Table 2 show the win-rates of our preference-trained models against state-of-the-art open-source models. Importantly, the base Aya 23 8B already achieves high win-rates against Gemma-1.1 (62.1%), Llama-3 (66.6%), and Mistral-v0.3 (69%) averaged across all 23 languages on Dolly open-ended generations. Preference-optimized Aya



(a) Avg. win-rates on ***23*** languages (b) Win-rates on ***English*** (c) Avg. win-rates on ***unseen*** langs.

Figure 2: DPO and RLOO results with an increasing number of languages in preference training data. We report win-rates for **(a)** the average of 23 languages, **(b)** only English, and **(c)** the average of unseen languages, reflecting the cross-lingual transfer.

models extend this lead further. Concretely, our best variant of RLOO leads to 69.5% (+7.4), 72.4% (+5.8), and 77.5 (+8.5) win-rates against Gemma, Llama-3, and Mistral respectively.

Win Rates Against Aya-23-8B Table 3 shows win-rate results for open-ended generations and summarization respectively. Win-rates are measured against Aya 23 8B. We report win-rates for both DPO and RLOO trained with different preference data mixtures. Due to the space constraints, summarization results are shown in Appendix (Table 8). Our best variant across experiments achieves 70.7% over the base Aya model.

Online optimization vs offline We find that RLOO (online RL) outperforms DPO (offline) across the board in multilingual performance. As shown in Table 3, RLOO achieves higher win-rates with all the preference data mixtures compared to DPO. For the 23-language mixture, RLOO achieves 54.0% win-rates whereas DPO reaches 47.0%.

Furthermore, RLOO achieves higher cross-lingual transfer with English-only and 5-language preference training in comparison to DPO. On languages unseen during training, RLOO achieves 3.4% higher win-rate compared to DPO (46.3% vs 42.9%) when training only on English and an 11.6% higher win-rate (54.9% vs 43.3) when training on the 5-language subset as shown in Table 5. This is inline with recent works (Tajwar et al., 2024; Tang et al., 2024) which suggest that online sampling outperforms fully offline methods. Additionally, the significant improvement in cross-lingual transfer with RLOO compared to DPO, suggests that online sampling also enables *generalization* beyond the distribution of the training prompts. This is complementary to findings of (Kirk et al., 2024b) which show that online RLHF enables better generalization than Supervised Fine-Tuning (SFT).

Increasing multilinguality in preference data improves winrates For a fixed dataset size of 50K pairwise preferences, we find that increasing the number of languages in training data improves the overall performance as shown in Figure 2a and Table 3. For DPO, win-rates against the base model on 23 languages increases from 43.3% to 47.0%.[6] For RLOO, this gain is most visible when the number of training languages is 5 where win-rate improves from 46.4% to 54.0% (+7.6 increase). Interestingly, in open-ended generations, using all 23 languages does not improve performance further for RLOO, however, in summarization, the 23-language training mixture increases win-rates compared to the 5-language subset (from 65.1% to 70.7%, Table 8).

English also benefits from multilingual training Our experiments also show that English can benefit from multilingual preference training and positive transfer, even when there are fewer total English examples in the training data. As shown in Figure 2b (and Table 4), our RLOO ML-23-50K variant outperforms our RLOO EN-1-50K variant (53.0% vs 47.5% win-rate) on English, despite being trained on 23 times less English data. However, English win-rates drop for DPO as the number of languages increases when there is a fixed budget of 50K examples, showing that DPO may be more prone to

[6]Win-rates that are under 50% do not correspond lower performance since our evaluation allows for *Tie*. To indicate if there is a performance gain, we also report the difference between win- and loss-rates (**ΔW-L**) in the results.

		Average 23 Languages		
		Win%	Loss%	ΔW-L%
BASE	GEMMA-1.1	62.1	29.4	32.7
	LLAMA-3	66.6	29.4	37.2
	MISTRAL-V0.3	69.0	26.8	42.2
DPO	GEMMA-1.1	67.7	27.1	40.6
	LLAMA-3	71.0	24.7	46.3
	MISTRAL-V0.3	74.7	21.8	52.9
RLOO	GEMMA-1.1	69.5	26.3	43.2
	LLAMA-3	72.4	24.0	48.4
	MISTRAL-V0.3	77.5	18.9	58.6

Table 2: Open-ended generation (Dolly) win-rates for the base Aya 23 8B, and DPO/RLOO preference optimized Aya 23 8B models against Gemma-1.1-7B-it, Meta-llama-3-8B-Instruct and Mistral-7B-Instruct-v0.3. We report average win-rates on 23 languages for the best ML-23-230K checkpoints for DPO and RLOO.

		Average 23 Languages		
		Win%	Loss%	ΔW-L%
DPO	EN-1	43.3	40.6	2.7
	ML-5	43.8	39.1	4.7
	ML-23	47.0	37.1	9.9
	ML-23*	50.2	39.0	11.2
RLOO	EN-1	46.4	38.9	7.5
	ML-5	54.4	35.8	18.6
	ML-23	54.0	38.0	16.0
	ML-23*	53.4	37.0	16.4

Table 3: Open-ended generation (Dolly) win-rates for DPO/RLOO preference optimized Aya models against the original Aya 23 8B. We report average win-rates on 23 languages for multiple training data mixtures: EN-1 (English Only), ML-5 (5 Languages), and ML-23 (23 Languages). All the data mixtures consist of 50K total training examples with the exception of ML-23*, which includes 230K total training examples. We report results for the best checkpoint across 2 epochs.

negative interference.

Cross-lingual transfer to unseen languages Preference training only with English, achieves performance gains for languages not seen in the training data as shown in Figure 2c (and Table 5). This gain (**ΔW-L**) is 2.0% for DPO and 7.3% for RLOO. Furthermore, using a 5-language subset (ML-5) significantly increases these gains to 3.8% and 19.4% for DPO and RLOO respectively. These results provide strong evidence of cross-lingual transfer in preference optimization, which is significantly more present after online training, with an increased degree of transfer facilitated by multilingual training data.

Role of data size and reward over-optimization For DPO, increasing the amount of multilingual data from approximately 2K to 10K per language in the 23-language mixture improves win-rates from 47.0% to 50.2% (Table 3). For RLOO, however, the same increase does not lead to an improvement. For all runs, the best checkpoint was the last one (epoch 2), except for the RLOO ML-23 230K run where we observed significant performance degradation after 0.5 epochs, which may be caused by reward model overoptimization (Gao et al., 2022b). Prior works have shown that low-resource languages can jailbreak LLMs (Yong et al., 2024) and it is likely that reward models, which usually are initialized from LLMs, likely share similar vulnerabilities. The degradation for the RLOO ML-23 230K run we observe may be caused by online optimization exploiting more languages and prompts where the reward model may be more prone to reward hacking, as this run includes all 23 languages and more prompts per language than other runs.

Is there a multilingual alignment tax? Post-training stages of LLMs including supervised fine-tuning and preference optimization have increasingly been torn between objectives: improving traditional discriminative benchmarks like HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021) and training LLMs to follow instructions, acquire conversational abilities, and be helpful and harmless (Askell et al., 2021). Recent work on multilingual instruction finetuning (Üstün et al., 2024) has found that improvements in open-ended generative tasks introduce trade-offs with discriminative tasks. However, this work only studies the tensions introduced by supervised instruction finetuning. More recent work (Tang et al., 2024) on preference training suggests that offline methods impart improved ability for discriminative tasks, whereas on-policy sampling improves generative quality. Here, we explore whether this holds in a multilingual setting and characterize the trade-off between discriminate and generative performance.

Table C shows multilingual benchmark results of our preference optimized Aya models and original Aya-23-8B. We follow Aryabumi et al. (2024) and evaluate models on the unseen tasks (XWinograd (Muennighoff et al., 2023b), XCOPA (Ponti et al., 2020), XStoryCloze (Lin et al., 2022)), mMMLU (Lai et al., 2023), and MGSM (Shi et al., 2023) using `eval-harness` framework (Gao et al., 2021).

We find that both DPO and RLOO are robust in

		English		
		Win%	Loss%	ΔW-L%
DPO	EN-1	52.0	33.5	18.5
	ML-5	50.5	28.5	22.0
	ML-23	44.5	36.5	8.0
	ML-23*	57.5	31.0	26.5
RLOO	EN-1	47.5	38.5	9.0
	ML-5	55.5	30.5	25.0
	ML-23	53.0	37.0	16.0
	ML-23*	53.0	35.0	18.0

Table 4: English Dolly win-rate results for DPO/RLOO preference optimized Aya 23 8B on multiple training data mixtures: EN-1 (English Only), ML-5 (5 Languages), ML-23 (23 Langauges). All runs are done with 50K total training examples with the exception of ML-23*, which is done with 230K total training examples. We report results for the best checkpoint across 2 epochs.

		Avg. Unseen Langs.		
		Win %	Loss %	ΔW-L%
EN-1	DPO	42.9	40.9	2.0
	RLOO	46.3	39.3	7.3
ML-5	DPO	43.3	39.5	3.8
	RLOO	54.9	35.5	19.4

Table 5: Win-rates for the 22 and 18 unseen languages that are not included in the training data for EN-1 and ML-5 respectively. We observe cross-lingual transfer from preference optimization, with an increased degree of transfer enhanced by multilingual training.

terms of their results in multilingual benchmarks, matching the performance of the base Aya 23 8B model. More specifically, multilingual preference optimization through both DPO and RLOO only lead to a slight drop in MGSM by an accuracy of 0.6 (36.6 vs 36.0) and 1.5 (36.6 vs 35.1) respectively. In contrast to recent work (Tang et al., 2024) in a monolingual setting, this shows that multilingual preference optimization can substantially improve generative performance as discussed in Section 4, while incurring minimal tax on common multilingual NLP tasks.

5 Related Work

Reinforcement Learning from Human Feedback (RLHF) RLHF has become the dominant paradigm for aligning LLMs to human preferences. Canonically, this involves training a reward model and using a reinforcement learning algorithm like PPO (Schulman et al., 2017a) or RLOO (Ahmadian et al., 2024) to optimize the LLM policy to maximize reward of online samples generated throughout training. (Ouyang et al., 2022b; Stiennon et al., 2020; Christiano et al., 2017a). There has been a plethora of work attempting to take the online inference aspect, and the optimization difficulties and complexities of RL that come with it, out of RLHF. The most prominent of these, is the family of methods base upon Direct Preference Optimization (DPO) (Rafailov et al., 2023), such as IPO (Azar et al., 2024), KTO (Ethayarajh et al., 2024), and SRPO (Choi et al., 2024). This family of methods directly fine-tunes an LLM to be impicitly consistent with collected preference data, forgoing the need for training a separate reward model.

Reinforcement Learning from AI Feedback (RLAIF) Collecting human feedback is often very expensive. Thus many recent works (Bai et al., 2022; Yuan et al., 2024), make use of feedback generated from AI, which can be LLMs which have been already optimized for alignment with human preferences with AI. Often these LLMs can be used to provide additional rankings, ratings, or natural language feedback, which can be used in subsequent RLHF training.

Preference Optimization for Multilingual LLMs There have been limited efforts on multilingual preference optimization to-date. (Lai et al., 2023) present a multilingual instruction tuning framework, where they preference train multilingual LLMs such as BLOOMZ (Muennighoff et al., 2023a) for 26 non-English languages with RLHF. They synthetically generated a preference dataset by translating an extended version of the Alpaca dataset (Taori et al., 2023), generating responses from their target LLM and ranking back-translated (into English) responses with ChatGPT.[7] In contrast to our work, they perform preference optimization for each language separately. However, due to their potentially low-quality dataset which heavily relies on translations, their resulting language-specific models are outperformed by other massively multilingual LLMs without preference optimization (Üstün et al., 2024; Aryabumi et al., 2024). Wu et al. (2024) study cross-lingual transfer in reward model (RM) training where they propose using preference data in one source language to train an RM for target language alignment. They show that RMs based on a multilingual base model exhibit zero-shot cross-lingual transfer consistently

[7] https://openai.com/blog/chatgpt/

across different languages. However, they do not experiment with using multiple source languages in training, which we show is crucial in the preference optimization both for offline optimization such as DPO and online RL methods such as RLOO.

6 Conclusion

Our work presents a comprehensive study on multilingual preference optimization. We show that the inclusion of multilingual data in preference optimization leads to significant improvements in multilingual performance over English-only preference optimization. This improvement scales both with the number of languages and the total number of examples included in the training data. Additionally, we show that preference training exhibits cross-lingual transfer, leading to significant gains in languages not present in the training data. We also find that using online preference optimization outperforms offline preference optimization, highlighting the importance of online samples during training.

As a result of our study, we expand the frontier of alignment techniques to 23 languages which cover half the world's population, by successfully preference-training an 8 Billion Aya 23 model that outperforms both the original Aya 23 8B base and widely used models including Gemma, Mistral, and Llama 3.

7 Limitations

A potential risk of relying on synthetic and translated datasets is the presence of particular cultural biases in model behavior. The prompts used in ShareGPT to seed the creation of the synthetic data over-index on contributions from the Global North or Western regions (Longpre et al., 2023). This could introduce a skew towards a narrow selection of cultural viewpoints.

Our preference-trained model covers 23 languages and improves performance relative to the closest open-source model. However, this is still only a tiny fraction of the world's linguistic diversity which encompasses 7000 languages. Furthermore, in this research we do not distinguish between dialects within the languages we cover, which are an important part of how language is used in practice (Zampieri et al., 2020; Wolfram, 1997; Brown et al., 2020; Lent et al., 2022; Blaschke et al., 2023; Falck et al., 2012). Future work, should aim to include more of the world's population and therefore languages.

Due to compute constraints, we are limited in our ability to run preference optimization experiments for larger models. Many of the runs we describe in this work for a single run can take 5 days to complete on a single 8 x H100 80GB GPU instance. Future work should explore the impact of scaling model size and further tune other hyperparameters for multilingual preference optimization.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *Preprint*, arXiv:2402.14740.

AI@Meta. 2024. Llama 3 model card.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment. *Preprint*, arXiv:2112.00861.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

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. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Hritik Bansal, John Dang, and Aditya Grover. 2024. Peering through preferences: Unraveling feedback acquisition for aligning large language models. In *The Twelfth International Conference on Learning Representations*.

Steven Bird. 2022. Local languages, third spaces, and other high-resource scenarios. pages 7817–7829.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Verena Blaschke, Hinrich Schuetze, and Barbara Plank. 2023. A survey of corpora for Germanic low-resource languages and dialects. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414, Tórshavn, Faroe Islands. University of Tartu Library.

Meriem Boubdir, Edward Kim, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. 2023. Which prompts make the difference? data prioritization for efficient human llm evaluation. *Preprint*, arXiv:2310.14424.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *Preprint*, arXiv:2307.15217.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2024. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *arXiv preprint arXiv:2404.08555*.

Eugene Choi, Arash Ahmadian, Matthieu Geist, Oilvier Pietquin, and Mohammad Gheshlaghi Azar. 2024. Self-improving robust preference optimization. *Preprint*, arXiv:2406.01660.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017a. Deep reinforcement learning from human preferences. *Preprint*, arXiv:1706.03741.

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

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models. *arXiv preprint arXiv:2310.06474*.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Preprint*, arXiv:2305.14387.

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef Genabith. 2022. Towards debiasing translation artifacts. In *Proceedings of the 2022*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3983–3991, Seattle, United States. Association for Computational Linguistics.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Oliver Falck, Stephan Heblich, Alfred Lameli, and Jens Südekum. 2012. Dialects, cultural identity, and economic exchange. *Journal of urban economics*, 72(2-3):225–239.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Leo Gao, John Schulman, and Jacob Hilton. 2022a. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*.

Leo Gao, John Schulman, and Jacob Hilton. 2022b. Scaling laws for reward model overoptimization. *Preprint*, arXiv:2210.10760.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. 2021. A framework for few-shot language model evaluation. *Version v0. 0.1. Sept*, page 8.

Gemini-Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan,

Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel

Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoff-

mann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma-Team. 2024. Gemma: Open models based on gemini research and technology.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *Preprint*, arXiv:2311.10702.

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.

Khyati Khandelwal, Manuel Tonneau, Andrew M. Bean, Hannah Rose Kirk, and Scott A. Hale. 2023. Casteist but not racist? quantifying disparities in large language model bias between india and the west. *ArXiv*, abs/2309.08573.

Md Tawkat Islam Khondaker, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Gptaraeval: A comprehensive evaluation of chatgpt on arabic nlp. *arXiv*, abs/2305.14976.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024a. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024b. Understanding the effects of rlhf on llm generalisation and diversity. *Preprint*, arXiv:2310.06452.

Wouter Kool, Herke van Hoof, and Max Welling. 2019. Buy 4 reinforce samples, get a baseline for free! In *DeepRLStructPred@ICLR*.

Hadas Kotek, Rikker Dockum, and David Q. Sun. 2023. Gender bias and stereotypes in large language models. *Proceedings of The ACM Collective Intelligence Conference*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and Jilin Chen. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. *arXiv*, abs/2310.16523.

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.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. Llms beyond english: Scaling the multilingual capability of llms with cross-lingual feedback. *Preprint*, arXiv:2406.01771.

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

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6439–6449, Marseille, France. European Language Resources Association.

Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. 2023a. Privacy in large language models: Attacks, defenses and future directions. *ArXiv*, abs/2310.10383.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023b. Making large language models better reasoners with step-aware verifier. *arXiv*, abs/2206.02336.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. 2023. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *Preprint*, arXiv:2310.16787.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Nils Lukas, A. Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-B'eguelin. 2023. Analyzing leakage of personally identifiable information in language models. *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023a. Crosslingual generalization through multitask finetuning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023b. Crosslingual generalization through multitask finetuning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Ranjita Naik, Varun Chandrasekaran, Mert Yuksekgonul, Hamid Palangi, and Besmira Nushi. 2023. Diversity of thought improves reasoning abilities of large language models. *arXiv*, abs/2310.07088.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv*, abs/2311.17035.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. 2022a. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022b. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2024. West-of-n: Synthetic preference generation for improved reward modeling. *Preprint*, arXiv:2401.12086.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. *Preprint*, arXiv:2403.19159.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. pages 2362–2376.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

Leonardo Ranaldi and Giulia Pucci. 2023. Does the English matter? elicit cross-lingual abilities of large language models. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 173–183, Singapore. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *Preprint*, arXiv:2310.10076.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. 2017a. Trust region policy optimization. *Preprint*, arXiv:1502.05477.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017b. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, Patrick Hall, et al. 2022. Towards a standard for identifying and managing bias in artificial intelligence. *NIST special publication*, 1270(10.6028).

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

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. *arXiv preprint arXiv:2402.06619*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. 2024. Preference fine-tuning of llms should leverage suboptimal, on-policy data. *Preprint*, arXiv:2404.14367.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, and Will Dabney. 2024. Understanding the performance gap between online and offline alignment algorithms. *Preprint*, arXiv:2405.08448.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin 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. *Preprint*, arXiv:2402.07827.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. 2023. On evaluating and mitigating gender biases in multilingual settings. *arXiv*, abs/2307.01503.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *Preprint*, arXiv:2404.18796.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. *Preprint*, arXiv:2312.08935.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. 2019. Characterizing and avoiding negative transfer. *Preprint*, arXiv:1811.09751.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256.

Walt Wolfram. 1997. Issues in dialect obsolescence: An introduction. *American speech*, 72(1):3–11.

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. 2024. Reuse your rewards: Reward model transfer for zero-shot cross-lingual alignment. *arXiv preprint arXiv:2404.12318*.

Nuo Xu, Jun Zhao, Can Zu, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Advancing translation preference modeling with rlhf: A step towards cost-effective solution. *arXiv preprint arXiv:2402.11525*.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2024. Low-resource languages jailbreak gpt-4. *Preprint*, arXiv:2310.02446.

Sicheng Yu, Qianru Sun, Hao Zhang, and Jing Jiang. 2022. Translate-train embracing translationese artifacts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 362–370, Dublin, Ireland. Association for Computational Linguistics.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *Preprint*, arXiv:2401.10020.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Background on DPO and RLOO

(1) Instruction fine-tuning (SFT) stage: A pre-trained LM is instruction-tuned using a dataset consisting of a given instruction prompt, and (typically) a human-written completion. The LM/policy is trained with a cross-entropy loss over the completion only. Often, the SFT model, denoted as π^{sft} is used to initialize both the reward model (for online RL optimization) and the RLHF policy model.

(2) Preference optimization stage: In this stage, the preference data such as rankings of model responses, are collected through humans or AI feedback. This data is then used to further fine-tune the SFT model (policy) to align the model with human feedback via the collected preferences data. Since collecting human feedback is often very expensive, many preference optimization methods train a separate reward model, on collected preference data to act as a proxy for human preferences, enabling for *online* preference feedback on LLM responses without requiring human intervention. Preference optimization can be performed in a number of ways:

Online Preference Optimization includes training a reward model, typically through binary classification, which is then used to provide online feedback in the optimization of the policy with the following objective:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(.|x)}[r_\phi(x,y) - \beta p_{KL}],$$
$$\text{with } p_{\text{KL}} = D_{\text{KL}}\pi_\theta(.|x)||\pi_{\text{ref}}(.|x)$$

where β is meant to control the distance from the initial policy π_{ref} during the optimization of reward $r_\theta(x,y)$ as proposed in (Stiennon et al., 2022). The KL-penalty p_{KL} is crucial as penalty-free optimization of the reward model leads to degradation in the coherence of the model.

Direct Preference Optimization (DPO; Rafailov et al., 2023) collects pairwise preferences (often over LLM responses) and fine-tunes the policy to beimplicitly consistent with the collected preference pairs by using the following loss:

$$-\log\sigma(\beta\log\frac{\pi_\theta(y_+|x)}{\pi_{\text{ref}}(y_+|x)} - \beta\log\frac{\pi_\theta(y_-|x)}{\pi_{\text{ref}}(y_-|x)})$$

Different from the online RL methods, DPO skips the reward modeling and enables preference optimization in an offline manner without requiring online samples from the policy during the training. At its core, DPO uses the analytical formulation of the canonical KL-controlled RLHF objective detailed in equation A and the assumption that preferences can be modeled by the Bradelly-Terry model (Bradley and Terry, 1952), to for-go reward modeling, simplifying the problem into a supervised style classification task.

While reinforcement learning approaches share the components above, techniques differ in formulating the reward and the sample-based loss. In this work, we use **REINFORCE-Leave-one-out (RLOO; Kool et al., 2019; Ahmadian et al., 2024)** estimator as the online preference optimization method as it is effective and more efficient than Proximal Policy Optimization (PPO) (Schulman et al., 2017b). RLOO is a multi-sample extension of REINFORCE (Williams, 1992), where multiple online generations are sampled from the policy per prompt which enables to reduce variance without requiring an additional network as opposed to PPO:

$$\frac{1}{k}\sum_{i=1}^{k}[R(y_{(i)},x) - \frac{1}{k-1}\sum_{j\neq i}R(y_{(j)},x)]$$
$$\nabla\log\pi(y_{(i)}|x) \text{ for } y_{(1)},...,y_{(k)} \overset{i.i.d}{\sim} \pi_\theta(.|x)$$

RLOO_k considers each $y_{(i)}$ individually and uses the remaining $k-1$ samples to create an unbiased estimate of the expected return for the prompt, akin to a *parameter-free* value-function, but estimated at each training step.

Model	Average 23 Languages Win	Tie	Loss
Base Aya 23 8B	49.9%	13.0%	37.1%
Gemma-1.1-7B	77.9%	3.5%	18.6%
Llama-3-8B	75.9%	4.4%	19.7%
Mistral-7B-v0.3	82.1%	3.3%	14.6%

Table 6: Win-rates for RLOO ML-23-230K model against other models with Claude 3.5 Sonnet as a judge averaged across 23 languages on Dolly. We observe similar trends to results Figure 1 where GPT-4-Turbo is the judge model.

B Additional Win-Rate Results

B.1 Claude 3.5 Sonnet as a Judge

Table 6 provides additional win-rate results with Claude 3.5 Sonnet as a Judge on Dolly averaged across 23 languages.

			English			Avg. 23 Langs.		
		Num Examples	Win%	Loss%	ΔW-L%	Win%	Loss%	ΔW-L%
RLOO vs DPO	EN-1	50K	37.5	46.5	-9.0	44.2	40.3	4.0
	ML-5	50K	50.5	40.0	10.5	51.4	38.8	12.6
	ML-23	50K	51.0	39.5	11.5	50.4	41.3	9.1
	ML-23	230K	50.0	35.0	15.0	47.4	41.1	6.2

Table 7: Direct win-rate comparisons for RLOO and DPO models on Dolly. RLOO models consistently outperform their DPO counterparts across all dataset splits.

		Average 15 Languages		
		Win%	Loss%	ΔW-L%
DPO	EN-1	52.5	41.6	10.9
	ML-5	50.5	43.4	7.1
	ML-23	53.0	40.9	12.1
	ML-23*	56.7	37.5	19.2
RLOO	EN-1	58.0	35.8	22.2
	ML-5	65.1	30.2	34.9
	ML-23	70.7	25.3	45.4
	ML-23*	68.9	26.7	42.2

Table 8: 15 language XLSum win-rate results for DPO/RLOO preference optimized Aya 23 8B on multiple training data mixtures: EN-1 (English Only), ML-5 (5 Languages), ML-23 (23 Langauges). All runs are done with 50K total training examples with the exception of ML-23*, which is done with 230K total training examples. We report results for the best checkpoint across 2 epochs.

B.2 RLOO vs DPO

To provide a head-to-head comparison between DPO and RLOO, Table 7 shows the win-rate evaluation between models trained with RLOO method with the models trained with DPO.

B.3 XLSum Summarization

Table 8 shows the win-rate scores of preference-trained models on 15 languages that are covered by our 23 language list. Win-rates are measured against the original Aya-23-8B model (Aryabumi et al., 2024). The average generation length for the base model, the best DPO, and the best RLOO models are 138, 234, and 171 tokens respectively. Length bias is a known property of DPO (Park et al., 2024) and can bias GPT-4 as an evaluator (Saito et al., 2023) accordingly. Because the base model and the RLOO model generation are similar in length, it is unlikely that the large gains in win-rate for the RLOO model against the base model are caused by GPT-4 as a judge preferring longer responses.

	Held out tasks (Accuracy %)			
Model	XCOPA	XSC	XWG	Avg
Base Aya 23 8B	59.8	62.3	80.7	67.6
DPO Aya 23 8B	59.9	62.6	80.7	67.7
RLOO Aya 23 8B	59.4	62.8	81.1	67.8

Table 9: Results for **discriminative unseen (held-out) task** evaluation. Results are reported as the zero-shot performance averaged across all languages of XCOPA, XStoryCloze, and XWinoGrad. DPO and RLOO checkpoints are for ML-23 230K runs

C Discriminative Benchmark Results

D Judge Prompt

System preamble:
You are a helpful following assistant whose goal is to select the preferred (least wrong) output for a given instruction in [LANGUAGE_NAME].

Prompt Template:
Which of the following answers is the best one for given instruction in [LANGUAGE_NAME]. A good answer should follow these rules:
1) It should be in [LANGUAGE_NAME]
2) It should answer the request in the instruction
3) It should be factually and semantically comprehensible
4) It should be grammatically correct and fluent.

Instruction: [INSTRUCTION]
Answer (A): [COMPLETION A]
Answer (B): [COMPLETION A]

FIRST provide a one-sentence comparison of the two answers, explaining which you prefer and why. SECOND, on a new line, state only ‘Answer (A)’ or ‘Answer (B)’ to indicate your choice. If the both answers are equally good or bad, state ‘TIE’. Your response should use the format:

Comparison: <one-sentence comparison and explanation>

Preferred: <‘Answer (A)’ or ‘Answer (B)’ or ‘TIE’>

	en	ar	de	es	fr	hi	id	it	nl	pt	ro	ru	uk	vi	zh	**Avg**
Base Aya 23 8B	54.6	45.1	50	50.9	51	39.7	48.8	50.7	49.7	50.8	49.9	47.8	46.8	46.5	47.1	48.2
DPO Aya 23 8B	54.9	45.7	50.0	51.1	51.3	40.0	49.0	51.2	49.8	51.1	49.9	48.0	47.0	46.8	47.6	48.5
RLOO Aya 23 8B	54.0	45.2	50.0	50.5	50.4	39.8	48.6	50.3	49.1	50.47	49.48	47.79	46.64	46.49	47.1	48.0

Table 10: **Multilingual MMLU** (`5-shot`) results for base, DPO, and RLOO Aya 23 models.

	de	en	es	fr	ja	ru	zh	**Avg**
Base Aya 23 8B	40.4	48.0	45.2	38.8	12.8	38.0	32.8	36.6
DPO Aya 23 8B	39.6	45.6	44.4	41.2	8.4	37.6	35.2	36.1
RLOO Aya 23 8B	39.6	46.4	38.4	39.6	14.0	34.8	33.2	35.1

Table 11: **Multilingual Grade School Math benchmark (MGSM)** results for . We use questions with answers followed by CoT prompt (`5-shot`) in the same language (`native_cot`) as the dataset and `strict-match` score as the evaluation metric.

	Win (%)	**Tie (%)**	**Loss (%)**	**ΔW-L (%)**
en	57.5	11.5	31.0	26.5
vi	54.5	10.5	35.0	19.5
tr	47.5	13.0	39.5	8.0
pt	51.0	7.0	42.0	9.0
de	50.0	10.0	40.0	10.0
ar	50.0	12.0	38.0	12.0
cs	45.0	11.5	43.5	1.5
el	48.0	6.5	45.5	2.5
es	39.5	10.5	50.0	-10.5
fa	51.5	10.5	38.0	13.5
fr	47.0	10.0	43.0	4.0
he	48.5	10.5	41.0	7.5
hi	57.5	10.5	32.0	25.5
id	52.5	13.0	34.5	18.0
it	50.5	11.0	38.5	12.0
ja	53.0	15.5	31.5	21.5
ko	51.0	12.5	36.5	14.5
nl	49.0	14.5	36.5	12.5
pl	50.5	10.0	39.5	11.0
ro	58.0	8.5	33.5	24.5
ru	46.5	8.0	45.5	1.0
uk	52.0	8.0	40.0	12.0
zh	45.0	13.0	42.0	3.0

Table 12: All language results for DPO ML-23-230K

E Full Language Set Win-Rates

We provide full win-rate results broken down for all 23 languages for the ML-23-230K DPO run in Table 12 and the the ML-23-230K RLOO run in Table 13

F Language List

We provide a list and description of all languages supported by Aya 23 8B which we use to perform multilingual evaluations in Table 14.

	Win (%)	**Tie (%)**	**Loss (%)**	**ΔW-L (%)**
en	53.0	12.0	35.0	18.0
vi	58.5	6.5	35.0	23.5
tr	54.5	10.0	35.5	19.0
pt	54.5	11.0	34.5	20.0
de	54.0	10.5	35.5	18.5
ar	49.5	12.5	38.0	11.5
cs	57.5	8.0	34.5	23.0
el	50.5	7.0	42.5	8.0
es	55.5	8.0	36.5	19.0
fa	56.0	12.0	32.0	24.0
fr	49.5	8.0	42.5	7.0
he	56.0	8.0	36.0	20.0
hi	62.0	12.0	26.0	36.0
id	49.5	9.5	41.0	8.5
it	51.0	10.0	39.0	12.0
ja	58.5	10.5	31.0	27.5
ko	50.5	9.5	40.0	10.5
nl	49.0	10.0	41.0	8.0
pl	52.5	5.5	42.0	10.5
ro	54.0	11.0	35.0	19.0
ru	51.5	6.0	42.5	9.0
uk	50.0	14.0	36.0	14.0
zh	50.5	10.0	39.5	11.0

Table 13: All language results for RLOO ML-23 230K

Code	Language	Script	Family	Subgrouping
ar	Arabic	Arabic	Afro-Asiatic	Semitic
cs	Czech	Latin	Indo-European	Balto-Slavic
de	German	Latin	Indo-European	Germanic
el	Greek	Greek	Indo-European	Graeco-Phrygian
en	English	Latin	Indo-European	Germanic
es	Spanish	Latin	Indo-European	Italic
fa	Persian	Arabic	Indo-European	Iranian
fr	French	Latin	Indo-European	Italic
he	Hebrew	Hebrew	Afro-Asiatic	Semitic
hi	Hindi	Devanagari	Indo-European	Indo-Aryan
id	Indonesian	Latin	Austronesian	Malayo-Polynesian
it	Italian	Latin	Indo-European	Italic
jp	Japanese	Japanese	Japonic	Japanesic
ko	Korean	Hangul	Koreanic	Korean
nl	Dutch	Latin	Indo-European	Germanic
pl	Polish	Latin	Indo-European	Balto-Slavic
pt	Portuguese	Latin	Indo-European	Italic
ro	Romanian	Latin	Indo-European	Italic
ru	Russian	Cyrillic	Indo-European	Balto-Slavic
tr	Turkish	Latin	Turkic	Common Turkic
uk	Ukrainian	Cyrillic	Indo-European	Balto-Slavic
vi	Vietnamese	Latin	Austroasiatic	Vietic
zh	Chinese	Han & Hant	Sino-Tibetan	Sinitic

Table 14: 23 languages supported in Aya 23 model (Aryabumi et al., 2024) with each language's script, family, and subgrouping