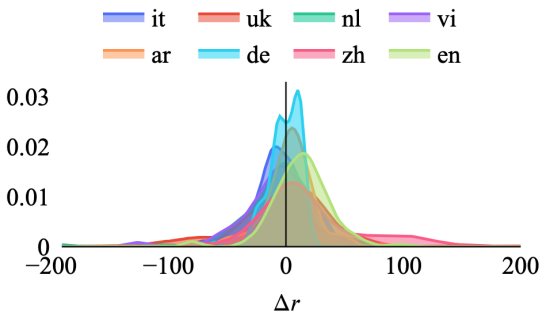
CAPO: Confidence Aware Preference Optimization Learning for Multilingual Preferences

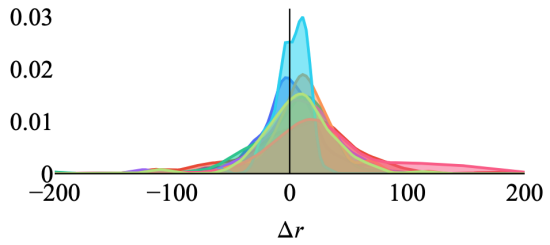
Rhitabrat Pokharel, Yufei Tao, Ameeta Agrawal
Department of Computer Science
Portland State University, USA
{pokharel, yutao, ameeta}@pdx.edu

Abstract

Preference optimization is a critical post-training technique used to align large language models (LLMs) with human preferences, typically by fine-tuning on ranked response pairs. While methods like Direct Preference Optimization (DPO) have proven effective in English, they often fail to generalize robustly to multilingual settings. We propose a simple yet effective alternative, Confidence-Aware Preference Optimization (CAPO), which replaces DPO's fixed treatment of preference pairs with a dynamic loss scaling mechanism based on a relative reward. By modulating the learning signal according to the confidence in each preference pair, CAPO enhances robustness to noisy or low-margin comparisons, typically encountered in multilingual text. Empirically, CAPO outperforms existing preference optimization baselines by at least 16% in reward accuracy, and improves alignment by widening the gap between preferred and dispreferred responses across languages.



(a) DPO



(b) CAPO

Figure 1: Density distribution of reward differences (Δr) across languages after DPO (top) and after applying our proposed confidence-aware preference optimization - CAPO (bottom). CAPO shifts the distributions towards higher Δr values, indicating improved separation between preferred and dispreferred responses.

1 Introduction

Preference optimization (PO) is a widely adopted post-training technique used to enhance the performance of large language models (LLMs) by aligning their outputs with human preferences. This is typically achieved by fine-tuning models using ranked responses or preference-based signals. While effective, most existing work in this area has been heavily centered on English (Rafailov et al., 2023; Meng et al., 2024a; Ethayarajh et al., 2024; Guo et al., 2024), with only recent efforts beginning to explore its application in languages other than English.

Multilingual alignment (Lai et al., 2023; Wu et al., 2024; Dang et al., 2024) increasingly adapts techniques such as PPO (Schulman et al., 2017), RLHF (Christiano et al., 2017), and especially DPO (Rafailov et al., 2023). However, existing methods still face challenges: DPO treats all preference margins equally (Yang et al., 2025b), RLOO relies on reward models, KTO (Ethayarajh et al., 2024) collapses rich feedback to binary labels, and DPL (Nath et al., 2025) uses fixed thresholds that may not generalize. Recent task-specific approaches continue to improve multilingual translation, reasoning, and safety (Xu et al., 2024b,a; She et al., 2024; Aakanksha et al., 2024).

We thus propose CAPO[1], Confidence Aware Preference Optimization, a simple yet effective enhancement to DPO that incorporates a Relative Reward Margin (RRM) into the optimization objective.

[1]The code is publicly available at https://github.com/PortNLP/CAPO.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1144–1156
December 20-24, 2025 ©2025 Association for Computational Linguistics

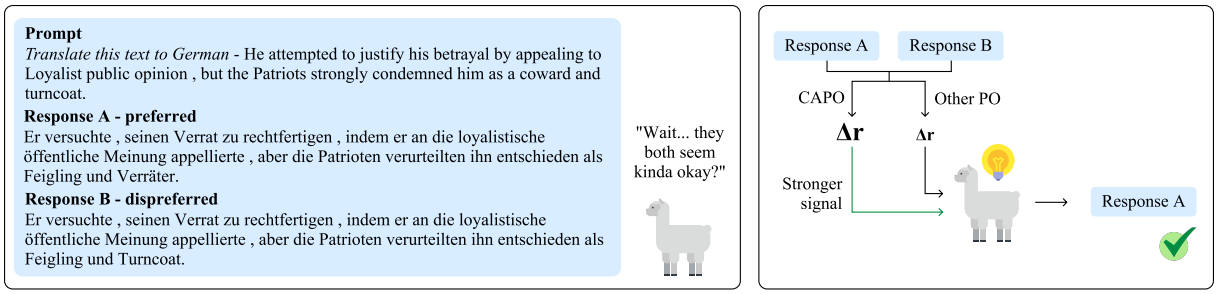


Figure 2: An example of a preference pair where both responses appear similarly plausible. CAPO leverages RRM to interpret such cases and provide a more informative learning signal. In other words, RRM boosts confidence in favor of the preferred response.

Unlike prior methods that treat all preference pairs equally, our approach dynamically adjusts the loss based on the relative difference in reward scores, allowing the model to calibrate its confidence during training.

As shown in Figure 1, this leads to a shift in the reward difference distributions towards the positive side across multiple languages. This is particularly important in multilingual settings, where language-specific inconsistencies and reward model uncertainty can make preference data noisier or less separable. Figure 2 illustrates how CAPO leverages RRM to improve ambiguous cases by adjusting the learning signal. By reweighting low-confidence or ambiguous examples and emphasizing clearer preference signals, RRM mitigates the risk of overfitting to hard multilingual cases and improves alignment robustness.

Our key contributions are as follows:

- We propose CAPO, a new PO technique designed for multilingual alignment. CAPO enhances DPO by dynamically adjusting loss based on confidence in preference pairs and also without requiring a reference policy.
- We demonstrate CAPO's effectiveness across three multilingual benchmarks, consistently outperforming relevant baselines.

2 CAPO: Confidence Aware Preference Optimization

2.1 Limitations of uniform weighting in DPO

Given a prompt x along with a preferred (winner) response y_w and a dispreferred (loser) response y_l, DPO optimizes the model by directly maximizing the likelihood difference between these responses. Formally, DPO optimizes the reward difference as:

$$\Delta r = \beta \log \frac{\pi(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}.$$

It encourages the model to favor the preferred response y_w over the dispreferred one y_l. While the loss dynamically adjusts its gradient based on the size of Δr, it does not distinguish whether the comparison is between two high-quality outputs or two low-quality, poorly aligned ones; it does so only in terms of absolute difference. It does not account for the *relative scale* of rewards across examples or languages. For instance, two preference pairs might yield the same reward difference, $\Delta r = r(y_w \mid x) - r(y_l \mid x) = 1$, but in one case the underlying rewards could be $r = [5, 4]$ and in another $r = [1.2, 0.2]$. The relative preference is much stronger in the later (1.2/0.2 > 5/4). This shows that absolute rewards fail to capture the strength of preference well when rewards are large. DPO treats both cases identically. This can become problematic in multilingual alignment, where reward distributions can vary significantly across languages. Empirically, we show that scaling the loss by RRM improves preference signal.

In a multilingual setting, tokenization rate varies from language to language. For example - for a similar sample in *en* and *ne*, the reward difference (Δr) between winning and losing will be bigger for *en* and smaller for *ne*. The bigger reward difference for *en* means small penalty and vice versa. DPO gives a different reward signal in this case, which can unfairly bias optimization toward languages with shorter tokenizations or higher per-token log-probs. Whereas in CAPO, RRM adjusts on a per-example basis to account for these tokenization rate differences. Instead of treating all reward differences uniformly, CAPO scales the loss by the confidence ($\sim$ RRM) derived from the ratio

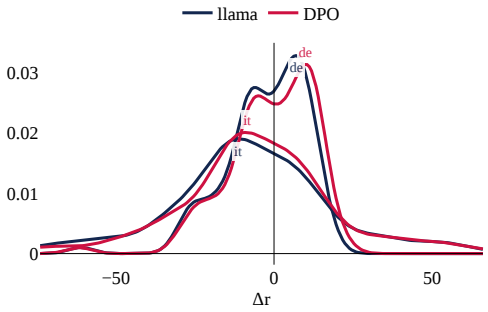


Figure 3: Shift in reward signal across *it* and *de* under llama and DPO. Although there is a shift of reward difference towards the right, there are a lot of samples for which the difference is negative.

of preferred to dispreferred log-probabilities. RRM is bigger for *ne* which means penalty close to *en*. This ensures that languages with inherently different tokenization structures receive reward signals that better reflect true preference strength rather than tokenization-induced disparities.

Consider an example shown in Figure 3. While DPO improves the reward margin in languages like *it* and *de*, it still produces a substantial number of preference pairs with negative Δr, indicating that the model fails to consistently prioritize the intended winning response. This suggests that the same reward signal can have very different implications depending on the language and context. Without a mechanism to calibrate these signals—such as through relative scaling—the model may overreact to weak signals or underreact to strong ones, leading to misaligned or unstable updates. A more robust alignment objective must therefore account not just for the direction of the preference, but also for how decisively that preference is expressed in each sample.

2.2 Relative Reward Margin

To improve multilingual preference optimization, we propose augmenting the standard DPO framework with a *Relative Reward Margin* (RRM). Given a prompt x with preferred and dispreferred responses y_w and y_l, respectively, and a policy model π with temperature β, the log-likelihood terms are:

$$\log \pi(y_w \mid x) = \text{preferred log-likelihood}, \quad (1)$$
$$\log \pi(y_l \mid x) = \text{dispreferred log-likelihood}. \quad (2)$$

Standard DPO is based on the Bradley-Terry (BT) objective, which models the probability of preferring y_w over y_l as a sigmoid function of their log-likelihood difference:

$$\mathcal{L}_{\text{BT}} = -\log \sigma\left(\beta\left[\log \pi(y_w \mid x) - \log \pi(y_l \mid x)\right]\right). \quad (3)$$

We incorporate RRM as a confidence-aware adjustment to this objective by scaling the log-ratio with a dynamic margin term, allowing the model to focus more on examples with clearer preference separation. The resulting modified objective $\mathcal{L}_{\text{CAPO}}$ encourages the policy to increase the relative margin between preferred and dispreferred responses.

$$\mathcal{L}_{\text{CAPO}}(\pi) = -\mathbb{E}_{(x,y_w,y_l)}\Big[\log \sigma\big(\beta(\log \pi(y_w \mid x) - \log \pi(y_l \mid x)) + \alpha \cdot \underbrace{\tfrac{\log \pi(y_w|x)}{\log \pi(y_l|x)}}_{\text{RRM}}\big)\Big] \quad (4)$$

where α is a tunable hyperparameter which adjusts the weight given to RRM. A higher α increases the emphasis on examples with a larger margin between preferred and dispreferred responses, while a lower α reduces this emphasis. Following Meng et al. (2024a), we eliminate the need for a reference model in our objective. We omit SimPO's length normalization since it is problematic in multilingual settings: tokenized lengths vary across languages, leading to unfair penalties for longer-tokenized languages (Tsvetkov and Kipnis, 2024).

The term RRM represents a ratio between preferred and dispreferred rewards and serves to adaptively modulate the loss based on the model's relative confidence in the given pair. When the preferred reward substantially exceeds the dispreferred one, RRM increases and the overall loss diminishes—signaling that the model is already well-aligned in this example. In contrast, when the reward margin is small, RRM downweights the learning signal to avoid overfitting on uncertain or noisy preference pairs. In other words, it dynamically calibrates the optimization signal based on the model's confidence in each preference pair, particularly beneficial in multilingual alignment, where preferences are harder to model due to linguistic variability. By emphasizing confident updates and softening gradients on ambiguous cases, our method encourages more stable convergence and robust generalization across diverse languages.

Method	Objective
DPO (Rafailov et al., 2023)	$-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)$
SimPO (Meng et al., 2024b)	$-\log\sigma\left(\frac{\beta}{\|y_w\|}\log\pi_\theta(y_w\|x)-\frac{\beta}{\|y_l\|}\log\pi_\theta(y_l\|x)-\gamma\right)$
DPONLL (Yang et al., 2025b)	$-\log\sigma\left(\beta\log\frac{\pi_\theta(y_w\|x)}{\pi_{\text{ref}}(y_w\|x)}-\beta\log\frac{\pi_\theta(y_l\|x)}{\pi_{\text{ref}}(y_l\|x)}\right)+\text{NLL}$
CAPO	$-\log\sigma\left(\beta\log\pi(y_w\mid x)-\beta\log\pi(y_l\mid x)+\alpha\cdot\frac{\log\pi(y_w\|x)}{\log\pi(y_l\|x)}\right)$

Table 1: Various preference optimization objectives given preference data $\mathcal{D}=(x,y_w,y_l)$, where x is an input and y_w and y_l are the winning and losing responses.

Lang. Direction	`prompt` (Source Text)	`dispreferred` (Machine translation)	`preferred` (MTPE text)	Dataset
en-tr	People may not anticipate that patience and understanding are also necessary for travellers returning home.	İnsanlar bu sabır tahmin edemeyeceğiniz ve anlayış da eve dönen yolcular için gereklidir.	İnsanlar, ülkesine dönenlere de sabır ve anlayış gösterilmesi gerektiğini tahmin edemeyebilir.	DIVEMT
en-de	He also begins an affair with Veronica Harrington, who bails him out.	Er beginnt auch eine Affäre mit Veronica Harrington, die ihn rettet.	Er beginnt auch eine Affäre mit Veronica Harrington, die ihn rettet.	MLQE-PE

Table 2: A sample each from the MTPE datasets.

3 Experimental Setup

3.1 Models and Implementation

We use `Llama-3.1-8B-Instruct` (llama) as the base model and apply parameter-efficient finetuning using LoRA (Hu et al., 2022) (additional LoRA settings can be found in Appendix A). The chosen and rejected samples are fed to the model in the form of a dialogue, where the examples are formatted using a custom Zephyr-style chat template (Tunstall et al., 2024). We conduct hyper parameter search to set the value of α, which is set to 2.0 based on maximum evaluation accuracy during training (more details in Section §4.4).

3.2 Baselines

Consistent with prior work (Dang et al., 2024), as our baselines, we consider base llama along with tuned versions of llama with DPO (Rafailov et al., 2023), SimPO (Meng et al., 2024a), and DPONLL (Yang et al., 2025b). For DPO setup, the reference policy is identical to the policy being optimized. We selected DPO as a relevant baseline over SFT, given prior studies demonstrating its superior performance (Wu et al., 2024; Yang et al., 2025b,a), and over RLHF due to its greater efficiency (Rafailov et al., 2023). We use the same LoRA setting for finetuning across all the objectives compared. Table 1 presents the optimization objectives evaluated in this work.

3.3 Preference Dataset for Training

Multiple methods for creating multilingual preference datasets have been explored (Dang et al., 2024; She et al., 2024; Yang et al., 2025b; Aakanksha et al., 2024); however, none of these datasets have been publicly released. As such, inspired by Berger et al. (2024), where they used Machine Translated Post Edited (MTPE) data for preference alignment on machine translation, we use MTPE data as our preference data. MTPE data provides a natural, human-grounded preference signal: the post-edited output reflects human judgments of correctness and fluency, while the corresponding raw machine translation often contains errors or stylistic flaws. This eliminates the need for synthetic preference labels and introduces realistic, linguistically diverse "hard examples".

Specifically, we repurpose two MTPE datasets—`DIVEMT` (Sarti et al., 2022) and `MLQE-PE` (Fomicheva et al., 2022)—into preference datasets, although they were not originally designed/used for this purpose. Together, they cover eight language directions with English as the source: *en-ar*, *en-it*, *en-nl*, *en-tr*, *en-uk*, *en-vi* (`DIVEMT`) and *en-de*, *en-zh* (`MLQE-PE`).

As shown in Table 2, each sample includes a source sentence, a machine translated output,

and a post-edited version, where the source sentence is the `prompt`, the post-edited version is the `preferred`, and the machine translation is the `dispreferred`. The prompt part is formatted using the following prompt template.

```
Translate this text from {src_lang} to
{tgt_lang}:

{src_lang}: {src_sent}
{tgt_lang}:
```

We filter the dataset to include only samples where the text length exceeds 50 characters, to make sure the model receives sufficient context. Data is balanced per translation direction, with 100 samples per direction. This gives us 800 samples for training and 800 for evaluation. All related analysis are done on the evaluation set. We perform validation on a held out 800 samples built in the same way as the training dataset for hyperparameter tuning.

3.4 Evaluation Benchmarks and Metrics

We consider three widely used multilingual benchmarks for evaluation (see implementation details in Appendix A).

Multilingual MT-Bench (Zheng et al., 2023) evaluates the capabilities and alignment of language models through two-turn dialogue prompts, covering open-ended tasks like writing, reasoning, and math. Models generate responses for each turn based on prompts provided by the benchmark. GPT-4-Turbo serves as the judge and assigns a score from 1 to 10 for each turn based on overall quality, including helpfulness, relevance, and fluency. We report the average of the two turns. We test on seven languages: the seen languages *[en, zh, it, and de]*, and the unseen languages *[fr, es, and jp]*, with 80 samples per language.

XLSum (Hasan et al., 2021) is a multilingual abstractive summarization dataset consisting of BBC news articles across a wide range of languages. Following prior work on multilingual preference optimization (Dang et al., 2024), we evaluate on 50 samples from each of six *seen* languages: *[en, zh, vi, ar, uk, tr]*, and three *unseen* languages: *[fr, ja, es]*. We report win rates with GPT4o as the judge model (and Rouge-L scores under Appendix C).

M-IFEval (Dussolle et al., 2025) is a multilingual instruction-following benchmark that covers 4 languages: one *seen* - *[en]*, and three *unseen* - *[fr, ja, es]*. It evaluates multilingual instruction-following by models using objective checks like string matching and rule-based evaluation rather than relying on the LLM-as-judge model. We report both strict (exact matches) and loose (acceptable variations) evaluation metrics.

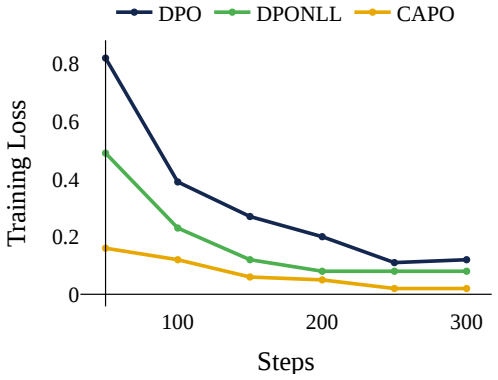


Figure 4: Training Loss vs. Steps for DPO and CAPO: CAPO demonstrates improved stability and convergence over DPO and DPONLL.

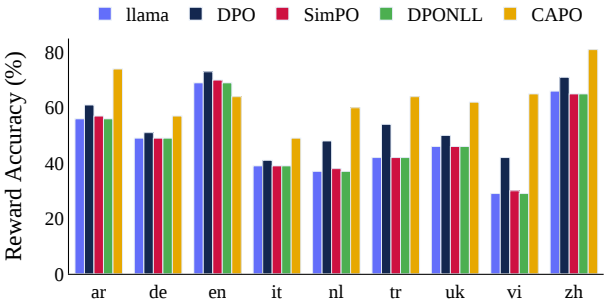


Figure 5: Comparison of reward accuracy between DPO and CAPO on the validation data. CAPO shows improved accuracy across all languages except in *en*.

4 Results and Analysis

The following subsections present key findings from our experiments with CAPO. We show that it improves training stability, sharpens reward signals, and increases the gap between preferred and dispreferred outputs. We also examine the effect of varying the α parameter. Finally, we report results on multilingual benchmarks.

4.1 CAPO makes training more stable.

We compare training curves for CAPO against two baselines: DPO and DPONLL (Yang et al., 2025b) in Figure 4. We observe clear differences in stability and convergence. DPO steadily improves reward but shows some fluctuations in later epochs.

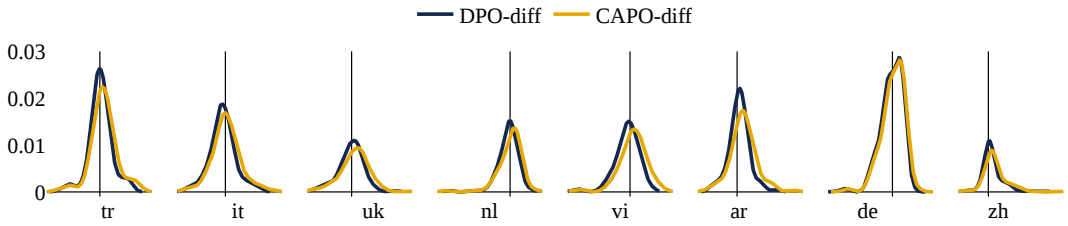


Figure 6: Reward difference distribution between DPO vs CAPO per language. CAPO consistently shifts the distribution toward higher reward differences.

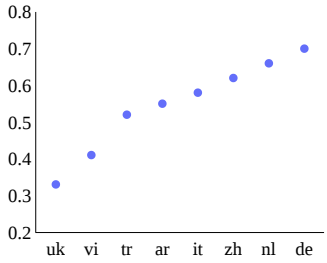


Figure 7: BLEU score between the preferred and dispreferred samples in the training data per language.

DPONLL is more stable, with smoother and more consistent reward growth. This stability comes from the added negative log-likelihood reweighting. CAPO shows the most stable and monotonic reward increase. It learns faster in early stages and then levels off gradually. This suggests that CAPO's confidence-aware design speeds up early training and supports stable alignment later.

We further investigate the lower training loss of CAPO compared to DPO and DPONLL. While DPO uses a contrastive term based on the log-sigmoid of log-probability differences, and DPONLL adds an unbounded negative log-likelihood (NLL) term that heavily penalizes low-probability preferred responses, CAPO avoids such penalties. Instead, CAPO combines the bounded ratio-based component (RRM) that contributes a smaller signal than NLL term. Since it adds only a mild adjustment (compared to NLL) on top of the reward difference, the overall loss remains lower. This makes CAPO's loss lower from the start.

4.2 The reward signal to the model gets better with CAPO.

Reward accuracy refers to how often a reward model assigns a higher score to the response that humans prefer in a pairwise comparison (more details under Appendix B). Figure 5 shows results across eight languages and 5 settings. On average, CAPO significantly outperforms DPO by 16% and DPONLL by 33%. It achieves the highest reward accuracy in all languages except English *en*. CAPO's poor performance in *en* may be due to using EN $\rightarrow$ other languages MTPE data, which might have introduced learned preferences that are subtly different from native English data. Vietnamese *vi* enjoys the most gain just like in the standard DPO. These results suggest that CAPO not only improves training stability but also leads to better reward modeling aligned with human intent. We attribute this to RRM, which helped deliver more reliable reward signals during optimization.

4.3 CAPO leads to increase in reward difference.

Figure 6 shows the distribution of reward differences between preferred and dispreferred responses under DPO and CAPO across eight languages. The reward difference reflects how confidently the model separates the preferred response from the rejected one. Compared to DPO (where the curves are more narrowly peaked around zero), CAPO's distributions often show a broader spread with a greater mass concentrated on the right side of zero on all languages. This indicates that under CAPO, the reward model assigns significantly higher scores to preferred responses more frequently, leading to stronger reward separation.

We conduct further analysis. Languages like *vi*, *ar*, and *uk* show larger shift towards the right which can be credited to the lower BLEU scores (i.e. lower similarity) in these languages (see Figure 7) between preferred and dispreferred samples. This hints that languages exhibit clearer preference signals benefit more from CAPO. Supporting this, we observe an inverse correlation ($r = -0.47$) between BLEU scores and the CAPO-vs-DPO reward shift across languages, computed by correlating BLEU scores with the difference in KDE-weighted means of reward distributions.

Prompt

Translate this text to German - He attempted to justify his betrayal by appealing to Loyalist public opinion , but the Patriots strongly condemned him as a coward and turncoat.

Response A - preferred

Er versuchte , seinen Verrat zu rechtfertigen , indem er an die loyalistische öffentliche Meinung appellierte , aber die Patrioten verurteilten ihn entschieden als Feigling und Verräter.

Response B - dispreferred

Er versuchte , seinen Verrat zu rechtfertigen , indem er an die loyalistische öffentliche Meinung appellierte , aber die Patrioten verurteilten ihn entschieden als Feigling und Turncoat.

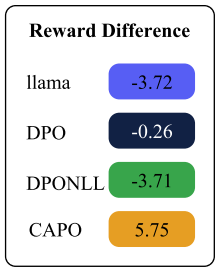


Figure 8: The reward difference shows how well each model distinguishes the responses. CAPO increases the gap between preferred and dispreferred responses the most.

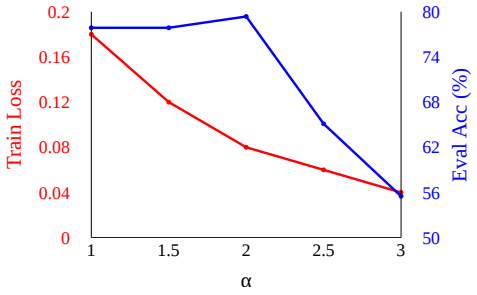


Figure 9: Impact of α on training loss and evaluation accuracy.

For instance, Figure 8 shows an example where the initial reward gap between the preferred (Response A) and dispreferred (Response B) is negative, with llama assigning -3.72. DPO slightly improves this to -0.26, while DPONLL remains similar at -3.71. CAPO significantly increases the reward gap to 5.75, showing a much stronger distinction in favor of the preferred response.

4.4 Impact of varying α.

We investigate the impact of the reweighting coefficient α in the $\mathcal{L}_{\text{CAPO}}$ objective, where α scales the log-ratio term that encourages stronger separation between preferred and rejected responses. As shown in Figure 9, increasing α consistently reduces training loss, indicating improved optimization. However, evaluation accuracy exhibits a tradeoff: it increases up to $\alpha = 2.0$ but drops sharply beyond that point. This suggests that while moderate reweighting strengthens preference alignment, excessive weighting of the RRM term leads to overreweighting and harms generalization. Overall, the results highlight a critical tradeoff between training loss minimization and evaluation performance when tuning α.

Lang.	llama	DPO	SimPO	CAPO
en	7.83	**8.12**	6.49	8.08
de	**7.43**	7.01	7.25	7.07
it	6.80	7.16	6.87	**7.39**
zh	6.55	6.60	6.87	**7.62**
fr	7.01	6.93	6.72	**7.40**
es	6.73	7.07	**7.72**	7.09
ja	6.85	6.71	6.75	**7.18**
Avg	7.03	7.09	6.95	**7.40**

Table 3: Results on multilingual MT-Bench benchmark.

4.5 Multilingual Benchmarks

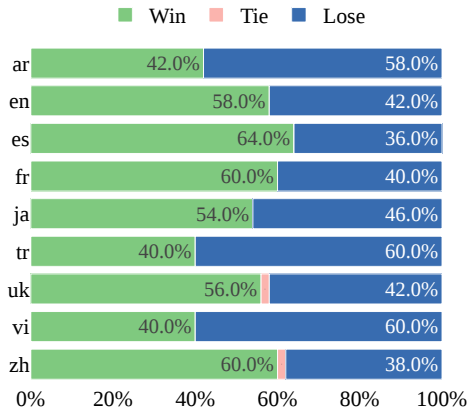
Next, we compare CAPO with the strongest baseline DPO in the following downstream benchmarks.

Multi-turn Conversation and Instruction Following (Multilingual MT-Bench) Table 3 presents results on the multilingual MT-Bench benchmark across seven languages. CAPO achieves the highest average score (7.40), outperforming DPO (7.09), SimPO (6.95), and the base llama model (7.03). Among the seen languages (*en*, *de*, *it*, *zh*), CAPO performs competitively, with notable improvements in *it* and *zh*, suggesting better generalization in training-aligned languages. For unseen languages (*fr*, *es*, *ja*), CAPO also shows consistent gains over other models except for *es*. This indicates that the confidence-aware objective used in CAPO improves not only in-domain alignment but also enhances cross-lingual generalization to languages not observed during training.

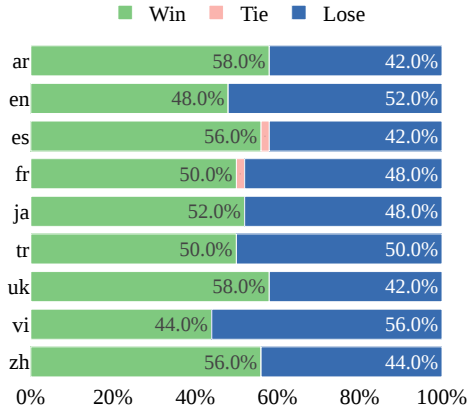
Summarization (XLSum) Figure 10 shows the pairwise win rates of CAPO compared to the base llama model (Figure 10a) and DPO (Figure 10b) across eight languages. Against llama, CAPO achieves substantial improvements in most languages, winning over 50% of the time in *en*, *zh*, *uk*, *fr*, *ja*, and especially *es* (64%). This demonstrates that preference optimization significantly en-

Text	llama	DPO	CAPO
The application is designed to stop using the computer's central processing unit (CPU) when it is left to run in the background. It means users can receive calls or instant messages without the risk of exhausting their laptop battery. The move may help Skype [...]	Microsoft has released a new version of Skype that can pause and resume activity without using the computer's central processing unit, allowing users to receive calls or messages without draining their laptop battery.	The software update for Skype allows the program to run in the background without consuming the computer's CPU power, thus saving battery life for laptops and tablets.	The new version of Skype will conserve laptop battery life by automatically suspending activity when the application is left to run in the background.

Table 4: An example of the summaries generated by different models and their head to head comparison against CAPO. Here, CAPO's answer is selected by the judge (GPT4o) both the times.



(a) Win rate of CAPO against llama.



(b) Win rate of CAPO against DPO.

Figure 10: Comparison of CAPO's win rates against llama and DPO for XLSum summaries. CAPO shows improved ability to generate better summaries across languages compared to the baselines. Standard deviations across runs are in Appendix E.

hances response quality over the base model. When compared to DPO, the margins are narrower but still favorable to CAPO in most cases. CAPO outperforms DPO in six out of eight languages, with strong gains in *ar*, *uk*, *zh*, and *es*, where win rates reach or exceed 56%. DPO outperforms CAPO is *en* and *vi*, with a 52% and 56% win rate respectively. These results indicate that CAPO offers robust multilingual improvements.

Lang.	Score Type	DPO	CAPO
en	Strict	0.46	**0.47**
	Loose	0.6	**0.63**
es	Strict	**0.56**	0.55
	Loose	0.68	**0.69**
fr	Strict	0.46	**0.48**
	Loose	0.63	**0.64**
ja	Strict	0.23	**0.29**
	Loose	0.41	**0.43**

Table 5: Comparison between the average scores (strict and loose) of M-IFEval using DPO and CAPO.

We further examined sample summaries generated by llama, DPO, and CAPO in en. Table 4 presents one example from each model for the same input text. A judge model compared the CAPO summary against those from llama and DPO. In both cases, the CAPO summary was rated better than the other two.

Instruction-Following (M-IFEval) Table 5 reports the accuracy of the response for the given prompts. Here again, we can see that CAPO consistently outperforms DPO under both the settings.

5 Related Work

Broadly, multilingual PO research can be categorized into general-purpose methods that align models across multiple languages using shared optimization mechanisms (Dang et al., 2024; Yang et al., 2025a,b), and task-specific strategies adapted to domains such as multilingual reasoning, safety alignment, text quality evaluation, and machine translation (She et al., 2024; Aakanksha et al., 2024; Xu et al., 2024b,a; Pokharel and Agrawal, 2025).

A major focus of recent work has been on cross-lingual alignment—developing techniques to transfer preferences learned in one language (typically

en) to others. Most prior approaches generate synthetic multilingual preference data via translation. Lai et al. (2023) translate both prompts and responses, Dang et al. (2024) translate only prompts, and Yang et al. (2025b) translate completions to favor dominant language outputs. While these cross-lingual strategies have enabled preference alignment in many languages, they often carry over linguistic artifacts and biases from the source language, which may not reflect the true human preference.

Regarding PO objectives, earlier studies have extended a range of PO techniques originally developed for *en*. Lai et al. (2023) and Wu et al. (2024) employ standard reinforcement learning methods such as Proximal Policy Optimization (PPO) (Schulman et al., 2017), with the latter placing emphasis on cross-lingual transfer during optimization. Cal-DPO (Xiao et al., 2024) calibrates implicit rewards to fixed targets using a reference model. Likewise, SimPO operates in a reference-free setting and applies length normalization, which is problematic in multilingual setting as discussed in Section 2.2. CPO (Xu et al., 2024c) uses a contrastive loss to distinguish preferred from dispreferred responses based on their log-probability differences and, similar to DPO, treats all pairs uniformly. Furthermore, CPO depends on a reference model, adding an additional layer of complexity.

Similarly, Dang et al. (2024) explore Reinforcement Learning with Optimal Outputs (RLOO) alongside DPO, while Yang et al. (2025b) apply DPO in conjunction with NLL of the preferred response. Although RLOO improves upon traditional RLHF approaches (Christiano et al., 2017) and PPO by reducing variance through multi-sample baselines, it still depends on a separately trained reward model, which introduces potential alignment gaps. Another PO method, Kahneman-Tversky Optimization (Ethayarajh et al., 2024), uses a prospect-theoretic objective to model human-like cognitive biases, but relies on binary desirability labels instead of preference pairs. Diverse Preference Learning (Nath et al., 2025) adjusts loss contributions based on the contrast between preferred and dispreferred samples, amplifying clear preferences and downweighting ambiguous ones. However, it relies on fixed thresholds and global weightings, which do not adapt to the reward dynamics of individual examples.

6 Conclusion

This paper introduces CAPO, a novel multilingual alignment objective that utilizes relative reward differences between preferred and dispreferred responses to guide the model's alignment. CAPO avoids reliance on a reward model and bypasses cross-lingual alignment methods that risk introducing translationese. Through empirical results on multilingual benchmarks, CAPO demonstrates significant improvements in aligning model outputs with human preferences across diverse languages, achieving 16–33% gains in reward accuracy. While we relied on existing MTPE data for training, future work should consider automatically generating high-quality MTPE data to expand language coverage and better reflect true human preferences.

Limitations

We conduct our experiments using LoRA-based fine-tuning with a small training set for each language. While full fine-tuning may yield different insights, our results demonstrate that even with efficient, lightweight training, CAPO achieves meaningful improvements. Additionally, there remain many languages that could not be studied in this work.

Ethical Consideration

While our study focuses on a small set of languages, other languages especially low-resource languages remain underexplored and could benefit significantly from improved alignment techniques.

Acknowledgement

We thank Dr. Suresh Singh, PortNLP Lab members, and the anonymous reviewers for their constructive feedback.

References

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

Nathaniel Berger, Stefan Riezler, Miriam Exel, and Matthias Huck. 2024. Post-edits are preferences too. In *Proceedings of the Ninth Conference on Machine*

Translation, pages 1289–1300, Miami, Florida, USA. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

John Dang, Arash Ahmadian, Kelly Marchisio, Julia Kreutzer, Ahmet Üstün, and Sara Hooker. 2024. RLHF can speak many languages: Unlocking multilingual preference optimization for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13134–13156, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Antoine Dussolle, A. Cardeña, Shota Sato, and Peter Devine. 2025. M-IFEval: Multilingual instruction-following evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6161–6176, Albuquerque, New Mexico. Association for Computational Linguistics.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Yiju Guo, Ganqu Cui, Lifan Yuan, Ning Ding, Zexu Sun, Bowen Sun, Huimin Chen, Ruobing Xie, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Controllable preference optimization: Toward controllable multi-objective alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1437–1454, Miami, Florida, USA. Association for Computational Linguistics.

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

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024a. SimPO: Simple preference optimization with a reference-free reward. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Abhijnan Nath, Andrey Volozin, Saumajit Saha, Albert Aristotle Nanda, Galina Grunin, Rahul Bhotika, and Nikhil Krishnaswamy. 2025. DPL: Diverse preference learning without a reference model. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3727–3747, Albuquerque, New Mexico. Association for Computational Linguistics.

Rhitabrat Pokharel and Ameeta Agrawal. 2025. Mtq-eval: Multilingual text quality evaluation for language models. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2025 (Findings)*. Association for Computational Linguistics.

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

Gabriele Sarti, Arianna Bisazza, Ana Guerberof-Arenas, and Antonio Toral. 2022. DivEMT: Neural machine translation post-editing effort across typologically diverse languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7795–7816, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Shuaijie She, Wei Zou, Shujian Huang, Wenhao Zhu, Xiang Liu, Xiang Geng, and Jiajun Chen. 2024. MAPO: Advancing multilingual reasoning through multilingual-alignment-as-preference optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10015–10027, Bangkok, Thailand. Association for Computational Linguistics.

Megh Thakkar, Quentin Fournier, Matthew Riemer, Pin-Yu Chen, Amal Zouaq, Payel Das, and Sarath Chandar. 2024. A deep dive into the trade-offs of parameter-efficient preference alignment techniques. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5732–5745, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Tsvetkov and Alon Kipnis. 2024. Information parity: Measuring and predicting the multilingual capabilities of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7971–7989, Miami, Florida, USA. Association for Computational Linguistics.

Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M Rush, and Thomas Wolf. 2024. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. 2024. Reuse your rewards: Reward model transfer for zero-shot cross-lingual alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1332–1353, Miami, Florida, USA. Association for Computational Linguistics.

Teng Xiao, Yige Yuan, Huaisheng Zhu, Mingxiao Li, and Vasant G Honavar. 2024. Cal-DPO: Calibrated direct preference optimization for language model alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. In *ICML*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024c. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025a. Implicit cross-lingual rewarding for efficient multilingual preference alignment. *arXiv preprint arXiv:2503.04647*.

Wen Yang, Junhong Wu, Chen Wang, Chengqing Zong, and Jiajun Zhang. 2025b. Language imbalance driven rewarding for multilingual self-improving. In *The Thirteenth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Appendix

A Implementation Details

Finetuning Following the configuration of Thakkar et al. (2024) on DPO finetuning, we set the LoRA rank to 16, alpha to 32, and apply a dropout rate of 0.05. Likewise, optimization is conducted using the Paged AdamW optimizer, paired with a cosine decay learning rate scheduler. Training is performed using the CAPO objective for 1 epoch, with a batch size of 16 and a learning rate of $1e{-}6$. We use a temperature parameter of $\beta = 0.1$, consistent with prior work.

MT-bench We use the multilingual MT-Bench framework [2] (Zheng et al., 2023), following the same setup as Yang et al. (2025b). The benchmark extends the original MT-Bench with multilingual

[2] https://github.com/lightblue-tech/multilingual-mt-bench

support in *en*, *de*, *it*, *zh*, *fr*, *es*, and *ja*, with 80 samples per language.

The base model evaluated in all experiments is `Llama 3.2 8B Instruct`. We compare three variants: the untuned base model, a DPO-finetuned model, and our proposed CAPO-finetuned model. All generations are produced using identical decoding settings, with a maximum of 1024 generated tokens. Sampling temperature is set to 0.7 by default. To ensure a fair comparison, the same evaluation pipeline is used across all model variants.

XLSum The evaluation prompt used to calculate win rate between two summaries is adapted from Yang et al. (2025b), which is shown in Figure 11. To mitigate order bias, the summaries were randomly shuffled.

B Reward Accuracy Calculation

As in SimPO, we compute *reward accuracy* directly from the trained model's own log-probabilities, without using a reference model. For each preference pair (x, y_w, y_l), we calculate the model scores as the log-likelihood of each response, following SimPO's formulation. Reward accuracy is then defined as the percentage of cases where the model assigns a higher score to the preferred response y_w than to the dispreferred response y_l.

C Multilingual Benchmarks

This section outlines the remaining analyses on the multilingual benchmarks.

XLSum - RougeL Figures 12 present Rouge-L scores on the XLSum dataset for summarization across multiple languages. CAPO achieves comparable average performance to DPO, both slightly outperforming the base LLaMA model. The per-language breakdown in Figure 12 reveals that CAPO shows more consistent or improved performance in several individual languages, except for *fr* and *ja*. The limited improvement may be due to the limitations of ROUGE-L, particularly when summaries are not too different at surface level. As suggested in prior work (Fabbri et al., 2021; Deutsch et al., 2021; He et al., 2023), automatic metrics ROUGE-L or BERTScore often fails to capture human preferences related to consistency, relevance, and fluency. Because ROUGE relies on lexical overlap, it is sensitive to exact word matches and cannot effectively detect semantic equivalence or paraphrasing. That is why we additionally report win rates to better compare against human

XLSum Win Rate Prompt

Which of the following answers is the best one for given instruction in [LANGUAGE]. A good answer should follow these rules:

1) It should be in [LANGUAGE]
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: Generate a one sentence summary of the text below in [LANGUAGE].

[TEXT]

Answer (A): [SUMMARY_A]
Answer (B): [SUMMARY_B]

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)' or 'TIE'.

Your response should use the format:
Comparison: <one-sentence comparison and explanation>
Preferred: <'Answer (A)' or 'Answer (B)' or 'TIE'>

Figure 11: Prompt used for win rate calculation on XLSum summaries.

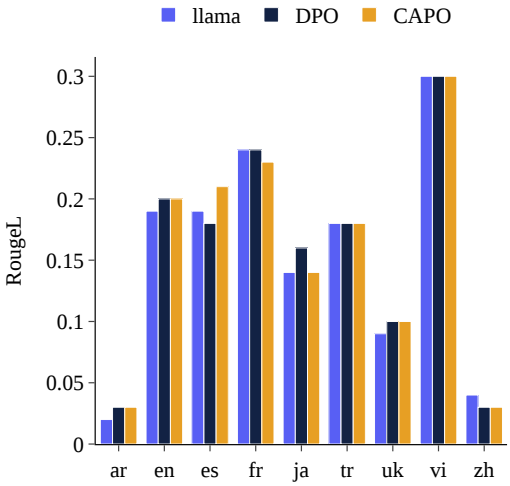


Figure 12: Average RougeL scores on XLSum dataset per language.

preferences as suggested in Wang et al. (2024).

D Language List

Here, we provide the list of languages along with their corresponding language codes.

Code	Language
Arabic	ar
Chinese	zh
Dutch	nl
English	en
French	fr
German	de
Italian	it
Japanese	ja
Spanish	es
Turkish	tr
Ukrainian	uk
Vietnamese	vi

Table 6: Language codes and their corresponding full language names.

E More Results

E.1 Reward Accuracy

Figure 13 presents the overall average reward accuracy results. CAPO outperforms DPO and DPONLL by 16% and 33% respectively.

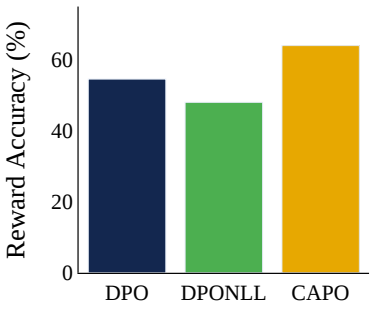


Figure 13: Comparison of average reward accuracy across different objectives on the validation data.

Lang.	SD (Win)	SD (Tie)	SD (Lose)
en	1.41	0	1.41
zh	5.66	2.83	2.83
vi	2.83	0	2.83
ar	1.41	0	1.41
uk	1.41	1.41	0
tr	0	0	0
fr	1.41	1.41	2.83
ja	2.83	0	2.83
es	4.24	0	4.24

Table 7: Standard deviation between two runs of XL-Sum win rates for CAPO vs llama.

Lang.	SD (Win)	SD (Tie)	SD (Lose)
en	1.41	0	1.41
zh	4.24	0	4.24
vi	4.24	0	4.24
ar	5.66	0	5.66
uk	1.41	0	1.41
tr	1.41	1.41	2.83
fr	1.41	0	1.41
ja	1.41	0	1.41
es	1.41	1.41	15.56

Table 8: Standard deviation between two runs of XL-Sum win rates for CAPO vs DPO.

E.2 XLSum

Tables 7 and 8 report the standard deviation between the two runs of XLSum. Except for es, the standard deviation remains low.