MAPO: Advancing Multilingual Reasoning through Multilingual-Alignment-as-Preference Optimization

Shuaijie She , Wei Zou , Shujian Huang* , Wenhao Zhu
Xiang Liu, Xiang Geng, Jiajun Chen
National Key Laboratory for Novel Software Technology, Nanjing University
{shesj,zouw}@smail.nju.edu.cn, huangsj@nju.edu.cn
{zhuwh,liuxiang,gx}@smail.nju.edu.cn, chenjj@nju.edu.cn

Abstract

Intuitively, reasoning abilities are considered language-agnostic. However, existing LLMs exhibit inconsistent reasoning abilities across different languages, e.g., reasoning in the dominant language like English is superior to other languages due to the imbalance of multilingual training data. To enhance reasoning abilities in non-dominant languages, we propose a Multilingual-Alignment-as-Preference Optimization framework (MAPO) to align the reasoning processes in other languages with the dominant language. Specifically, we harness an off-the-shelf translation model for the consistency between answers in non-dominant and dominant languages, which we adopt as the preference for optimization, e.g., Direct Preference Optimization (DPO) or Proximal Policy Optimization (PPO). Experiments show that MAPO stably achieves significant improvements in the multilingual reasoning of various models on all three benchmarks (MSVAMP +16.2%, MGSM +6.1%, and MNumGLUESub +13.3%), with improved reasoning consistency across languages [1].

1 Introduction

The reasoning ability of large-scale models has attracted much attention (Cobbe et al., 2021; Wang et al., 2022; Wei et al., 2023; Yu et al., 2023). Though we consider reasoning to be language-agnostic, existing studies (Chen et al., 2023) show that due to the imbalance of pre-training and fine-tuning data across languages, the mathematical reasoning ability of existing large-scale models varies across languages, e.g., English, is far superior to that of the other languages.

To improve the reasoning ability in other languages, Chen et al. (2023) translated the English reasoning processes into other languages for supervised fine-tuning (Ouyang et al., 2022, SFT). Although SFT brings in preliminary capability for multilingual reasoning, we argue that two problems hinder further improvement.

Firstly, the annotation for the reasoning process is expensive to obtain even for the dominant language, and the reasoning processes involve complex mathematical reasoning, which may result in translation errors (Chen et al., 2023). As a result, the translated reasoning annotation for non-dominant languages is limited in both scale and quality. Without sufficient and diverse data, the results of supervised training are limited. It may also suffer from generalization issues in versatile task scenarios (Zheng et al., 2023; Ouyang et al., 2022), indicating potential difficulty in adapting to the out-of-domain test set.

More importantly, although supervised training with translated reasoning processes improves the reasoning ability for almost all trained languages, this strategy only fills in the missing reasoning annotation for non-dominant languages that originate from the dominant language. Thus the inherent gap between dominant and non-dominant languages is hardly narrowed.

Different from existing attempts, we propose to use the reasoning ability of the dominant language as the director for improving non-dominant languages. As the reasoning process is critical for obtaining the correct result, the multilingual reasoning ability may be improved by encouraging the reasoning in non-dominant languages to be similar to that in the dominant language. Therefore, we propose a **M**ultilingual-**A**lignment-as-**P**reference **O**ptimization (MAPO) framework by aligning the reasoning process of non-dominant languages to the dominant. Notably, MAPO exploits the strong reasoning ability in the dominant language and requires no annotation for the reasoning process.

More specifically, MAPO consists of two stages:

*Corresponding author

[1]The project is available at https://github.com/NJUNLP/MAPO

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10015–10027
August 11-16, 2024

preference estimation via multilingual alignment and preference optimization. During preference estimation, the reasoning processes to the same question are sampled from the LLM in both dominant and non-dominant languages. A well-trained, off-the-shelf translation model is employed to yield the translation probability between the reasoning in dominant and non-dominant languages. Since higher translation probability indicates a more consistent reasoning aligned with the dominant language, the corresponding reasoning in the non-dominant language is considered better and shall be promoted. During preference optimization, Proximal Policy Optimization (Schulman et al., 2017, PPO) and Direct Preference Optimization (Rafailov et al., 2023, DPO) are adopted to optimize the previously estimated preference. So the LLMs are trained to reason in the non-dominant languages as they do in the dominant language. We also conduct Iterative DPO for further preference optimization.

Experiments are conducted on three challenging multilingual reasoning test sets, namely MSVAMP (Chen et al., 2023), MGSM (Shi et al., 2022), and MNumGLUESub constructed from NumGLUE (Mishra et al., 2022), each covering 10 languages. MAPO achieves accuracy improvements of up to 16.2%, 6.1%, and 13.3% on the three benchmarks, respectively, reaching state-of-the-art performance on 7B models, thereby demonstrating the effectiveness of our method. Among them, the improvement on the out-of-domain dataset MSVAMP (+16.2%) shows that, by enhancing multilingual reasoning consistency, MAPO improves the multilingual reasoning capabilities of the model in a generalizable manner.

2 Preliminary

2.1 Multilingual Reasoning

A straightforward measurement to evaluate the efficacy of large language models (LLMs) lies in their proficiency in tackling complex reasoning, e.g., their performance in solving mathematical reasoning. Recent work (Wei et al., 2023; Wang et al., 2022) has verified a substantial improvement when LLMs are guided through a step-by-step reasoning process, instead of conducting a direct answer.

Some research has introduced mathematical reasoning datasets in the form of application problems, such as GSM8K (Cobbe et al., 2021), NumGLUE (Mishra et al., 2022), and SVAMP (Patel et al., 2021). To evaluate the multilingual reasoning capabilities of LLMs, Shi et al., 2022 propose MGSM by manually translating 250 samples of the GSM8K test set from English to other languages. Subsequent researches focus on enhancing the multilingual reasoning of LLMs. Chen et al., 2023 translate the GSM8K training data into other languages for supervised fine-tuning (SFT), which improves the model's multilingual reasoning capabilities. However, SFT suffers from data scarcity and catastrophic forgetting. Its out-of-domain generalization is also hard to guarantee (Zheng et al., 2023). Multilingual reasoning via LLMs remains an open challenge.

2.2 Preference Optimization

SFT maximizes the likelihood of annotated outputs and equips models with preliminary capabilities. However, models still exhibit various issues after SFT. Some researchers have further adjusted model behaviors and enhanced model capabilities through preference optimization.

RLHF (Zheng et al., 2023; Bai et al., 2022b) further rectifies these model behaviors via reinforcement learning by preference. RLHF introduces a reward model $r_\theta(x, y)$ given input x with its corresponding output y that captures the preference nuance from the human feedback. Then, $r_\theta(x, y)$ scores arbitrary LLM outputs y given input x for iterative policy rectifications during proximal policy optimization (Schulman et al., 2017, PPO). The tuning is guided by Eq 1 to maximize the expected rewards of the LLM policy π_θ with the minimum deviation from the SFT policy:

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}_{(x,y)\sim D_\pi}[r_\theta(x, y) - \gamma \log \frac{\pi_\theta(y|x)}{\pi_{\text{SFT}}(y|x)}], \tag{1}$$

where π_{SFT} is the original LLM via SFT, γ is a hyperparameter that constrains policy updates.

Though RLHF via PPO is effective in adapting LLMs to versatile human preferences, it involves four sub-models, making the training complex and costly. DPO (Rafailov et al., 2023) proposes to distill a referential SFT policy π_{ref} by polarizing the preference. DPO tuning involves a pair of outputs (y_w, y_l) in Eq 2:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta\log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right], \tag{2}$$

where y_w is favored over y_l, and β is a hyperparameter.

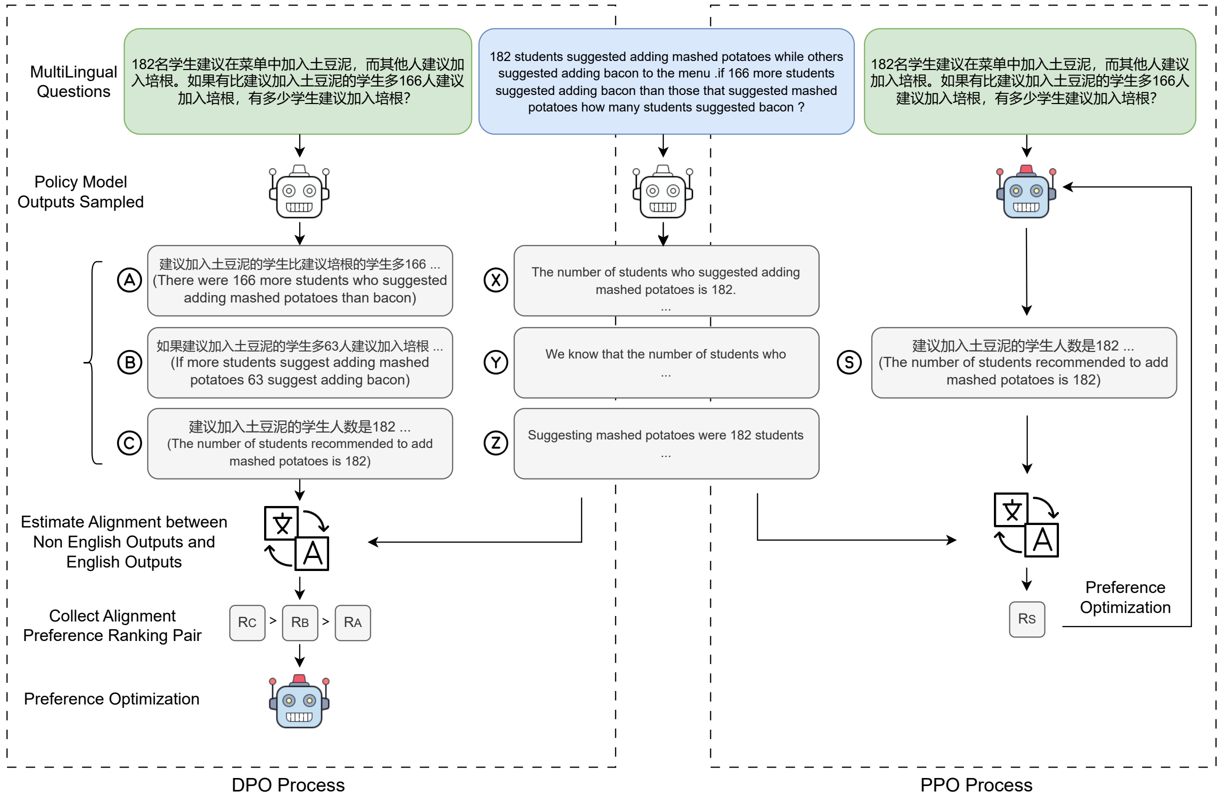

Figure 1: Illustration of our alignment framework. For brevity, we only show three sampling results and simplified optimization processes of DPO and PPO. The green and blue colors represent the same problem in Chinese and English, respectively. The white robots represent the original policy model and the colored robots represent the policy model with parameters updated through preference optimization.

3 Method

3.1 Preference Estimation

Intuitively, reasoning ability is language-agnostic. However, LLM reasoning varies across different languages, where we consider a dominant language to provide a better reference in reasoning for lesser languages. Therefore, a straightforward approach is to align reasoning in non-dominant languages to a dominant language. That is, we can rectify LLM's output reasoning in non-dominant languages by favoring answers that are more aligned with the dominant language.

We simply adopt a well-trained multilingual translation model to calculate the 'alignment' scores between answers in non-dominant and dominant languages as a preference. Since the translation model is optimized on a large scale parallel data in source and target languages $(Y, \bar{Y})$ by maximizing conditional generation probability in Eq 3:

$$\arg\max_{\theta} P(\bar{Y}|Y;\theta), \tag{3}$$

with more aligned target language $\bar{Y}$ to Y, the higher the conditional generation probability $P(\bar{Y}|Y)$. Then, we input the answers in non-dominant languages and force-decode a corresponding answer $\bar{Y}$ in the dominant language for the 'alignment' score. Correspondingly, answers with higher 'alignment' scores in non-dominant languages are preferred during the following preference optimizations.

3.2 Preference Optimization

With 'alignment' scores as a preference, we adopt two state-of-the-art preference optimizations: PPO and DPO, to optimize the alignment scores of reasoning in non-dominant languages given that in a dominant language. To make life easier, since English is dominant in both reasoning and translation, we adopt English as the dominant language in the rest of the paper.

3.2.1 Optimization with PPO

In our setup, the 'alignment' score $P(\bar{Y}|Y)$ given query x depicts the preference as the reward model $r_\theta(x, Y)$ do in PPO, thus we directly adopt the multilingual machine translation model as the reward model. During each optimization, we collect a batch of non-English outputs $\{Y_i\}$ with an English output $\bar{Y}$, respectively. PPO then maximizes the expected alignment score within the batch by Eq 4:

$$\mathcal{L}_{\text{PPO}} = \mathbb{E}_i[P(\bar{Y}|Y_i, x) - \gamma \log \frac{\pi_\theta(Y_i|x)}{\pi_{\text{SFT}}(Y_i|x)}]. \quad (4)$$

3.2.2 Optimization with DPO

DPO involves an input pair (Y_w, Y_l), where Y_w is favored over Y_l. Correspondingly, we collect n outputs within one non-English language, making $\binom{n}{2}$ pairs of (Y_w, Y_l) ranked by the alignment scores given the same English answer $\bar{Y}$. We additionally copy and freeze the original supervised fine-tuned LLM as the reference policy π_{ref} in Eq 2. Finally, our DPO is optimized by Eq 5:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,\bar{Y},Y_w,Y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(Y_w \mid x)}{\pi_{\text{ref}}(Y_w \mid x)} - \beta \log \frac{\pi_\theta(Y_l \mid x)}{\pi_{\text{ref}}(Y_l \mid x)}\right)\right]. \quad (5)$$

3.2.3 Iterative DPO

Recent studies such as LLaMA2 (Touvron et al., 2023) and Claude (Bai et al., 2022a) suggest that preference data updated through multiple iterative rounds is beneficial for preference optimization. Therefore, we further conduct iterative DPO to optimize reasoning alignments across languages. Specifically, we will further train π_{θ_i} on the preference data sampled by itself, which yields an updated model $\pi_{\theta_{(i+1)}}$. Subsequently, we continue the iteration by preference data sampled given $\pi_{\theta_{(i+1)}}$ for ongoing updates.

4 Experiment

4.1 Experiment Setup

Base Models Our experiments[2] include four base models. **MathOctopus-7/13B** is fine-tuned from LLaMA2-7/13B with the multilingual reasoning processes MGSM8KInstruct[3] (Chen et al., 2023). Yu et al., 2023 proposed MetaMath-7/13B and MetaMath-Mistral-7B, which have stronger English reasoning abilities. We fine-tune these models with MGSM8KInstruct and get **MetaMathOctopus-7/13B** and **MistralMathOctopus-7B**.

[2]More training details in Appendix A

[3]GSM8K translated to nine non-English languages.

Dataset	Type	Number
MGSM	In-Domain	2,500
MNumGLUESub	In-Domain	5,530
MSVAMP	Out-of-Domain	10,000

Table 1: Statistics of three benchmarks. “Type" indicates whether its corresponding training set was used during the training. “ Number” refers to the total number of samples in the test data.

Preference Estimation To obtain the mathematical problems for preference optimization, we selected tasks 1, 4, and 8 from the eight tasks in NumGLUE (Mishra et al., 2022), which is a multi-task arithmetic reasoning benchmark, and translate the questions into the same nine languages as in MGSM8KInstruct, resulting the dataset MNumGLUESub. MNumGLUESub with only 1700 problems. When constructing the preference dataset for DPO, we use the corresponding base model for sampling and calculate the alignment feedback using NLLB-600M-distilled [4].

Preference Optimization We experimented with both PPO and DPO. For simplicity, we report the results of the third iteration of DPO by default. For detailed results of PPO-LoRA and each round in Iterative DPO, please refer to Appendix B.

Evaluation Datasets We utilize three challenging benchmarks: MSVAMP, MGSM (Shi et al., 2022), and MNumGLUESub (statistics are shown in Table 1). Among them, MSVAMP, serving as an Out-of-domain test set that does not participate in the training, is used to analyze the model’s robustness and generalization ability. MGSM is the testset corresponding to MGSM8KInstruct, on which the base models are trained. MNumGLUESub is the testset corresponding to the data for preference estimation and optimization.

Evaluation Metric

- **Accuracy**: We use the accuracy of problem-solving to measure the model’s reasoning ability, with a higher accuracy representing stronger reasoning ability.
- **PPL (PPL-based Alignment Score)**: We input the target non-English answer and apply NLLB-600M-distilled to get the perplexity of the given English answer. Less perplexity indicates better alignment between the reasoning

[4]https://huggingface.co/facebook/nllb-200-distilled-600M

Model	Bn	Th	Sw	Ja	Zh	Ru	De	Es	Fr	En	Avg
GPT-3.5-Turbo	33.7	42.9	46.2	45.6	46.2	48.9	50.4	50.7	50.9	50.5	46.6
MAmmoTH 7B†	4.3	6.3	4.2	26.7	26.8	33.7	39.6	42.9	39.9	45.1	26.3
WizardMath 7B†	16.1	17.0	10.3	37.9	36.3	37.4	39.2	44.8	37.7	48.5	32.5
MetaMath 7B	14.8	17.7	14.8	51.8	54.4	59.4	59.6	63.2	62.0	64.8	46.2
MathOctopus 7B	27.7	35.9	39.4	41.6	42.7	44.2	44.0	45.1	45.3	46.4	41.2
+ m-RFT	37.9	46.4	46.4	49.6	50.8	50.4	50.7	51.6	53.4	49.4	48.7
+ MAPO-DPO(ours)	48.8	55.2	56.0	60.3	58.8	58.3	58.1	59.7	60.8	58.4	57.4
MetaMathOctopus 7B	36.1	47.5	49.4	51.3	54.5	53.6	56.6	60.0	57.2	64.2	53.0
+ m-RFT	44.8	54.2	56.2	58.4	57.7	56.2	59.2	59.3	57.8	63.1	56.7
+ MAPO-DPO(ours)	50.1	61.6	61.7	65.9	65.7	64.8	68.4	68.5	68.6	71.6	64.7
MistralMathOctopus 7B	47.0	52.6	54.4	58.8	60.2	59.8	62.1	60.8	60.4	74.4	59.0
+ m-RFT	57.4	63.5	60.8	65.2	69.8	67.4	67.2	69.1	68.0	71.3	66.0
+ MAPO-DPO(ours)	**62.9**	**71.3**	**71.4**	**73.7**	**76.0**	**74.9**	**77.8**	**78.1**	**79.0**	**81.1**	**74.6**
MAmmoTH 13B†	5.0	13.7	12.9	42.2	47.7	50.7	52.3	53.9	53.8	53.4	38.6
WizardMath 13B†	13.7	16.3	12.5	29.5	37.0	43.8	48.7	49.4	49.4	56.3	35.7
MetaMath 13B	14.6	15.7	17.4	57.0	56.6	63.7	67.3	65.9	64.7	67.7	49.1
MathOctopus 13B	35.2	46.8	42.8	43.2	48.8	47.6	44.4	48.0	48.4	53.2	45.8
+ m-RFT	43.4	50	52.1	54.9	55.4	57.1	59.2	56.4	59.5	55.2	54.3
+ MAPO-DPO(ours)	51.8	58.9	56.4	60.4	58.8	62.1	63.5	62.0	61.7	65.0	60.1
MetaMathOctopus 13B	41.6	52.1	50.9	57.3	53.1	59.1	60.1	61.1	60.8	66.8	56.3
+ m-RFT	48.1	59.6	61.4	60.5	58.9	61.0	62.7	65.3	64.3	65.4	60.7
+ MAPO-DPO(ours)	**54.7**	**64.7**	**62.9**	**69.0**	**68.2**	**68.2**	**69.5**	**70.6**	**71.3**	**70.5**	**67.0**

Table 2: Model Performances on MSVAMP test set. "Avg" represents the average performance in ten languages and bold text denotes the best results within the same size. Results marked with † come from Chen et al., 2023.

processes.

- **ACR (Answer Consistency Ratio)** : Let m denote the set of questions answered correctly in English, and n denote those answered correctly in non-English. ACR is then calculated as: ACR $= \frac{|m \cap n|}{|n|}$. Higher ACR indicates a greater degree of overlap in reasoning capabilities between non-English and English languages.

Baselines The selected base models are already strong baselines, which have been fine-tuned on MGSM8KInstruct. Moreover, motivated by Rejection sampling Fine-Tuning (RFT) (Yuan et al., 2023), we employ another strong baseline **m-RFT**, where the solutions that yield correct answers during sampling are used to further fine-tune the base model. To alleviate catastrophic forgetting, we fine-tune the model for only one epoch with a minor learning rate (1e-5). For comparison, we also incorporated other recent LLaMA2-base models: **MAmmoTH** (Yue et al., 2023) is developed by utilizing diverse datasets for math instruction, while **WizardMath** (Luo et al., 2023) employs Reinforcement Learning from Evol-Instruct (RLEIF). **MetaMath** 7B (Yu et al., 2023) is fine-tuned from the strongest English reasoning dataset, MetaMathQA.

5 Experiment Results

5.1 Preference Optimization Improves Multilingual Reasoning Effectively

Experimental results[5] in Table 2,3,8 consistently demonstrate that our method has effectively enhanced the reasoning capabilities of the various base models and achieved state-of-the-art performance. More specifically, the improvement is most impressive on the out-of-domain dataset MSVAMP, where we achieved an average accuracy improvement of 16.2% and 14.7% MathOctopus7B and MathOctopus13B, respectively. Even for the most powerful 7B model MistralMathOctopus 7B, our method can further boost its performance to an impressive 74.6%. We also observe significant improvements in MGSM and MNumGLUESub.

From the perspective of languages, most languages have improvement after alignment and it is more significant in some low-resource languages that previously under-performed. Take MathOctopus 7B as an example, our method has increased the accuracy on MSVAMP for Bengali, Thai, Swahili, and Japanese by 21.1%, 19.3%, 16.6%, and 18.7%,

[5]we facilitate a multilingual reasoning leaderboard: https://huggingface.co/spaces/kevinpro/Open-Multilingual-Reasoning-Leaderboard

Model	Bn	Th	Sw	Ja	Zh	Ru	De	Es	Fr	En	Avg
GPT-3.5-Turbo	31.2	38.0	40.0	36.0	44.0	43.2	46.0	47.2	41.6	54.4	42.2
MAmmoTH 7B†	3.6	4.8	2.4	10.8	17.2	26.0	33.2	32.4	32.8	49.6	21.3
WizardMath 7B†	2.0	4.0	3.4	24.0	22.4	30.8	30.4	34.8	30.4	47.6	23.0
MetaMath 7B	6.4	4.0	3.2	39.2	38.8	47.2	56.8	58.0	52.8	63.2	37.0
MathOctopus 7B	29.2	33.6	36.4	35.2	39.2	38.8	44.8	42.4	43.2	52.0	39.5
+ m-RFT	25.6	31.2	28.8	34.0	39.2	36.0	34.8	34.4	36.4	43.2	34.4
+ MAPO-DPO(ours)	30.8	38.0	37.6	45.2	47.2	42.0	45.2	43.2	40.8	45.6	41.6
MetaMathOctopus 7B	25.6	42.8	36.4	40.0	46.4	46.8	49.6	54.4	46.4	66.4	45.5
+ m-RFT	23.2	33.6	34.0	34.0	47.2	43.2	45.6	47.6	44.8	62.8	41.6
+ MAPO-DPO(ours)	36.0	44.8	44.8	47.6	55.2	53.6	53.6	56.8	52.4	70.8	51.6
MistralMathOctopus 7B	44.0	54.4	50.4	55.6	59.2	58.8	62.4	62.0	56.8	76.0	58.0
+ m-RFT	41.2	46.8	46.8	48.4	57.2	62.8	61.6	59.2	57.6	72.0	55.4
+ MAPO-DPO(ours)	**55.2**	**60.4**	**61.6**	**58.0**	**74.0**	**70.8**	**67.6**	**74.0**	**69.2**	**82.0**	**67.3**
MAmmoTH 13B†	3.6	5.2	1.6	19.2	31.2	36.8	45.6	50.0	39.6	56.4	28.9
WizardMath 13B†	6.4	5.6	5.6	22.0	28.0	34.4	40.4	45.6	42.0	52.8	28.3
MetaMath 13B	11.6	6.4	7.6	42.8	49.2	63.6	64.8	65.2	65.2	67.2	44.4
MathOctopus 13B	42.4	39.2	44.8	38.8	49.6	45.2	48.4	53.6	43.2	54.8	46.0
+ m-RFT	29.6	30.8	34.4	36.4	40.4	39.2	42.0	42.8	40.4	48.0	38.4
+ MAPO-DPO(ours)	38.8	46.8	42.0	47.6	53.6	49.2	52.0	54.4	46.4	54.0	48.5
MetaMathOctopus 13B	34.4	42.8	41.6	49.2	52.8	54.4	54.4	59.2	53.6	71.6	51.4
+ m-RFT	22.8	29.6	30.4	35.2	39.2	40.0	43.6	43.6	41.2	59.2	38.5
+ MAPO-DPO(ours)	**44.8**	**47.6**	**55.2**	**56.0**	**59.6**	**59.2**	**59.2**	**63.6**	**62.8**	**71.6**	**58.0**

Table 3: Model Performances on MGSM test set. "Avg" represents the average performance in ten languages and bold text denotes the best results within the same model size. Results marked with † come from Chen et al., 2023.

respectively.

Surprisingly, our preference optimization dataset does not contain English, but the accuracy of English has also improved. After alignment, MetaMathOctopus even surpassed the English mathematical reasoning models on English questions across three datasets. We believe this is primarily due to the alignment facilitating a more consistent understanding of reasoning across different languages, which contributes to the enhancement of general reasoning capabilities.

Meanwhile, our approach has also achieved stable and impressive improvements on the MetaMathOctopus 7B which has stronger reasoning capabilities, propelling it past the ChatGPT and the larger scale model MetaMathOctopus13B, demonstrating the robustness and potential of our framework.

5.2 Alignment is the Key to Enhanced Multilingual Reasoning

We adopt PPL and ACR to evaluate the degree of alignment in reasoning processes and final answers between other languages and English, respectively.

Regarding the alignment of the reasoning process, as shown in Figure 2, our method effectively improves the consistency of the reasoning process, particularly for languages where the base model had poorer alignment, such as Bengali, Thai, and Japanese. This proves that the model can show more similar reasoning thinking to English on non-English after alignment.

Additionally, we notice that alignment also contributes to more consistent final answers. A higher ACR indicates that there is a greater overlap between the questions answered correctly in non-English and those answered correctly in English. From the results in Figure 3, our alignment framework has greatly increased the ACR for each language. This means that the performance gains in Table 2 stem from the parts that intersect with English. These observations demonstrate that our method effectively aligns the model's non-English reasoning processes with English, thereby enhancing reasoning abilities in non-English languages.

5.3 Generalizable Multilingual Reasoning

As indicated in Table 2, on the out-of-domain test set, MAPO achieved a 16.2% and 11.7% increase in accuracy, which is significantly higher than the 7.5% and 3.7% achieved by m-RFT. Surprisingly, while MAPO effectively improves the multilingual reasoning performance on both two datasets, it also

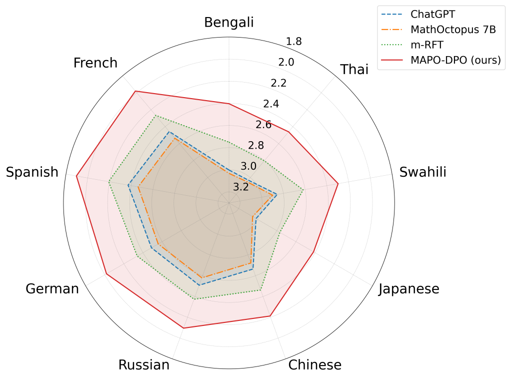

Figure 2: PPL across nine languages on MSVAMP. Lower PPL indicates higher consistency in reasoning processes.

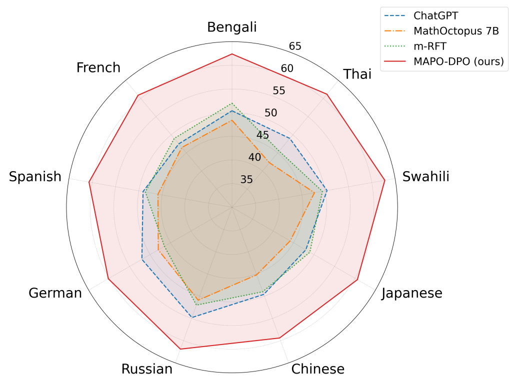

Figure 3: ACR across nine languages on MSVAMP. Higher ACR indicates higher answer consistency.

	MGSM	**MNumGLUESub**
Acc. on MSVAMP	31.6	54.9
Self-BLEU(En)	0.81	0.53
Self-BLEU(Non-En)	0.84	0.60

Table 4: Comparison between MGSM8KInstruct (MGSM) and MNumGLUESub. We report the average Self-BLEU scores for the sampled reasoning in English and non-English.

System	MSVAMP	MGSM
MetaMathOctopus7B	53.0	45.5
+ Ours & NLLB 600M	61.1	48.9
+ Ours & MBART-MMT-600M	59.9	49.6
+ Ours & M2M-1.2B	61.4	49.3

Table 5: Average accuracy on ten languages on MSVAMP and MGSM. Note that we only conduct one iteration MAPO-DPO with different translation models.

improves the performance on MGSM. In contrast, m-RFT exhibited a 5.1% and 3.9% performance degradation.

We believe that this is mainly because the model directly learns the given labels during SFT. The labels inevitably involve data-specific attributes, which are difficult to generalize. Conversely, our method does not offer the model with "correct answers", but guides the model to generate outputs we prefer, which is more effective in allowing the model to learn multilingual reasoning with better generalization.

6 Analysis

6.1 Preference Estimating on a Dataset Different from SFT

We conduct experiments with preference optimization using the MGSM8KInstruct dataset instead of the MNumGLUESub. Table 4 presents the average accuracy over 10 languages on MSVAMP. The results indicate that preference optimization based on the SFT dataset hurts performance on the out-of-domain test set MSVAMP. We suggest that SFT forces the model to stick to the oracle reasoning process, which hurts generation diversity. Thus, the policy model struggles to follow the desirable alignment preferences and easily falls into the trap of overfitting.

For further analysis, We randomly sample 1000 reasoning processes of LLMs from the MGSM8KInstruct and MNumGLUESub, then calculated the Self-BLEU (Zhu et al., 2018) on the sampled reasoning process within each problem-solving. The results (Table 4) suggest that on the MGSM dataset, LLMs tend to generate more consistent reasoning compared with MNumGLUESub, which confirms the diversity degradation.

6.2 Robustness over Different Translation Models

To verify the robustness of our framework across various translation models, we conduct experiments using MBART-MMT-600M [6] and M2M-1.2B [7], which have the different model architectures, scales, and training datasets. The results, presented in Table 5, demonstrate that our framework consistently achieves performance improvements. The performance is even slightly enhanced when with M2M-1.2B. These experimental results conclusively confirm the robustness of our framework.

[6] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[7] https://huggingface.co/facebook/m2m100_1.2B

English Question	182 students suggested adding mashed potatoes while others suggested adding bacon to the menu. if 166 more students suggested adding bacon than those that suggested mashed potatoes how many students suggested bacon ?
Chinese Question	182名学生建议在菜单中加入土豆泥，而其他人建议加入培根。如果有比建议加入土豆泥的学生多166人建议加入培根，那么有多少学生建议加入培根?
English Solution	The number of students who suggested mashed potatoes is 182. The number of students who suggested bacon is 182+166=348. ✓
Chinese Solution	**[Before Alignment]** 建议加入土豆泥的学生比建议加入培根的学生多166人，所以两边减去166得到差值为182 - 166 = 16，因此，建议加入培根的学生人数为16。(There are 166 more students recommending mashed potatoes than bacon so subtract 166 from both sides to get a difference of 182-166 = 16. Therefore, the number of students recommended to join Bacon is 16.) ✗ **[PPL]:** 2.65 **[After Alignment]** 建议加入土豆泥的学生人数是182。建议加入培根的学生人数是182 + 166 = 348。(The number of students who suggested mashed potatoes is 182. The number of students who suggested bacon is 182+166=348.) ✓ **[PPL]:** 0.97

Table 6: An example from the MSVAMP test set. The example clearly demonstrates that, through our preference optimization method, the model successfully corrects erroneous reasoning processes and generates reasoning processes that is more consistent with English.

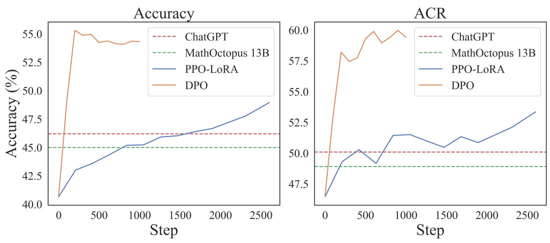

Figure 4: Average accuracy on non-English math problems and ACR versus preference optimization step on MSVAMP. MathOctopus 13B and ChatGPT are selected for comparison.

In addition, we have conducted experiments on translation models of different scales, and the results also demonstrate the robustness of our method. For more details, please refer to the Appendix D.

6.3 Reasoning Alignment during Preference Optimization

To better analyze the optimization preference process, we tested the checkpoints of the DPO and PPO-LoRA on MSVAMP and visualized the accuracy and ACR respectively in Figure 4.

The results show that the optimization of DPO is quite efficient. DPO enabled the MathOctopus 7B model to achieve higher accuracy and ACR than both ChatGPT and MathOctopus 13B within just 100 steps. As the optimization continues, the accuracy remains relatively stable while the degree of alignment still shows an increasing trend.

Meanwhile, although the optimization speed of PPO is slower, it take fewer than 1000 steps to surpass MathOctopus 13B. As the model continued to explore and optimize the preferences, it achieved superior multilingual reasoning performance compared to ChatGPT at 1600 steps.

The experimental results demonstrate that the model continuously aligns non-English reasoning with English reasoning, effectively enhancing its multilingual reasoning capabilities. For PPO-LoRA, although optimization is halted around 2500 steps due to the limitations on computational resources, we notice that growth in multilingual reasoning consistency and ability continues even beyond 2500 steps, indicating the potential of our preference optimization framework.

6.4 Alignment Reasoning for Improvements

The example in Table 6 directly illustrates how our framework can improve reasoning ability by aligning reasoning in other languages to English. For the given English problem, the basic model correctly analyzes that an addition should be conducted to find the number of students recommended to join Bacon. However, when given the same problem in Chinese, LLM misjudges the relationship of variables and writes a significantly different reasoning process. This directly proves that although the baseline model has been fine-tuned with multilingual reasoning data, its reasoning and thoughts are inconsistent, whereas non-English reasoning is more

System	MSVAMP	MGSM
MetaMathOctopus7B	53.0	45.5
+ Ours & En-as-Dominant	61.1	48.9
+ Ours & Es-as-Domiant	59.9	47.7

Table 7: Average accuracy on ten languages on MSVAMP and MGSM. Note that we only conduct one iteration MAPO-DPO with different dominant language.

prone to errors. After alignment, the reasoning thoughts are more similar to the English answer, and the reasoning thoughts are also corrected.

6.5 Flexible Choice of the Dominant Language

Due to the relatively strong translation and reasoning capabilities of English, our main experimental section reported results aligned with English. In addition, there are other high-resource languages such as Spanish, German, etc., which also possess relatively strong reasoning abilities. Intuitively, our method can also take these languages as the dominant language and enhance multilingual reasoning capabilities. Therefore, we chose Spanish as the target alignment language and repeated the aforementioned experiments. The experimental results, as shown in Table 7, indicate that aligning with Spanish can also improve the model's multilingual reasoning capabilities, although the extent of improvement is slightly less than that of English. This demonstrates the flexibility and robustness of our method; when English resources are not available or other languages perform better, our method can still effectively enhance the model's capabilities.

7 Conclusion

In this paper, we propose MAPO, a novel multilingual alignment-as-preference optimization framework, enhancing reasoning ability in non-dominant languages by aligning them with dominant languages. Experimental results demonstrate that our framework achieves significant improvements on various base models across all three benchmarks, especially with a notable 16.2% increase in average accuracy on the out-of-domain datasets MSVAMP. The analysis confirms that enhancing alignment through our method is the key to improvements in multilingual reasoning capabilities.

8 Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding author. This work is supported by the National Science Foundation of China (No. 62376116, 62176120), the Liaoning Provincial Research Foundation for Basic Research (No. 2022-KF-26-02), the research project of Nanjing University-China Mobile Joint Institute.

Limitation

Similar to previous work on preference optimization, our method necessitates a policy model that has preliminary multilingual reasoning capabilities through SFT. Meanwhile, due to limitations in computational resources, our experiments are confined to aligning models of 7B and 13B size and exploring two preference optimization algorithms, PPO and DPO. Should resources permit, we aim to extend our exploration to models of 70B sizes and examine the performance of a broader spectrum of preference optimization algorithms.

References

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. 2022b. Training a helpful and harmless assistant with reinforcement learning from human feedback. CoRR, abs/2204.05862.

Nuo Chen, Zinan Zheng, Ning Wu, Linjun Shou, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. 2023. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. arXiv preprint arXiv:2310.20246.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. CoRR, abs/2110.14168.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, and Dongmei Zhang. 2023. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. CoRR, abs/2308.09583.

Swaroop Mishra, Arindam Mitra, Neeraj Varshney, Bhavdeep Sachdeva, Peter Clark, Chitta Baral, and Ashwin Kalyan. 2022. Numglue: A suite of fundamental yet challenging mathematical reasoning tasks. arXiv preprint arXiv:2204.05660.

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. 2022. Training language models to follow instructions with human feedback.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems?

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. CoRR, abs/2305.18290.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. CoRR, abs/1707.06347.

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. 2022. Language models are multilingual chain-of-thought reasoners.

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. arXiv preprint arXiv: 2307.09288.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. CoRR, abs/2309.12284.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models. arXiv preprint arXiv:2308.01825.

Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2023. Mammoth: Building math generalist models through hybrid instruction tuning. CoRR, abs/2309.05653.

Rui Zheng, Shihan Dou, Songyang Gao, Yuan Hua, Wei Shen, Binghai Wang, Yan Liu, Senjie Jin, Qin Liu, Yuhao Zhou, Limao Xiong, Lu Chen, Zhiheng Xi, Nuo Xu, Wenbin Lai, Minghao Zhu, Cheng Chang, Zhangyue Yin, Rongxiang Weng, Wensen Cheng, Haoran Huang, Tianxiang Sun, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, and Xuanjing Huang. 2023. Secrets of RLHF in large language models part I: PPO. CoRR, abs/2307.04964.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. CoRR, abs/1802.01886.

A Experiment Details

Our code is primarily based on the trl [8], with some minor modifications made. The modified code will also be made available at our project.

Prompt: During the sampling, training, and testing phases, we consistently use the same prompt as MathOctopus (Chen et al., 2023).

LoRA: For experiments using LoRA, such as PPO LoRA, we optimize the [q_proj, v_proj, o_proj] modules. The values of r and alpha are set to 128 and 256, respectively.

DPO: We employ a learning rate of 1e-6, with β set at 0.1, and a warmup step count of 100. For models like Mistral that originally adopted a lower learning rate (5e-6) during their SFT phase, we will reduce the learning rate to 2e-7. The batch size is configured to 128. The optimization process is capped at a maximum of 1000 steps, and we save the checkpoint that corresponds to the lowest loss on the validation set. The training takes around 4 hours on 8 H100 GPUs.

PPO: We have configured a learning rate of 2e-5, with a batch size of 64 and 2 ppo epochs. We adopt the AdamW optimizer to improve the stability of the optimization, setting the epsilon value to 1e-5. Additionally, we have implemented a linear learning rate warm-up technique for the first 150 steps. All other hyperparameters are set to the default values provided by the trl library. We set the maximum optimization steps to 2600 and report the results at this checkpoint.

B Supplemental Experiment Results

To verify the robustness of our method, we conducted experiments on different preference optimization algorithms. Due to the limited computational resources, we optimized the PPO algorithm using LoRA. The detailed experimental results of PPO LoRA and each round in Iterative DPO are shown in Table 9.

Experiments demonstrate that our framework achieves an effective improvement in multilingual reasoning capabilities based on both PPO and DPO. Despite the limited computational resource, where PPO only updated a part of the parameters with merely 2600 steps, it has already brought impressive performance enhancements on all three datasets. Additionally, with the increasing rounds of DPO, the model exhibited progressively stronger multilingual reasoning abilities that have not yet reached its limit, revealing the potential of our approach.

[8] https://github.com/huggingface/trl

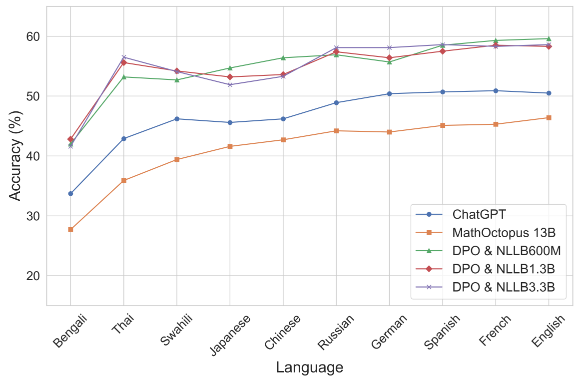

Figure 5: Accuracy across ten languages on MSVAMP after training MathOctopus 7B on preference datasets constructed using translation models of different scales.

C Used Scientific Artifacts

Below are the scientific artifacts used in our work. For the sake of ethics, our use of these artifacts is consistent with their intended use.

- *Transformers (Apache-2.0 license)*, a framework to facilitate downloading and training state-of-the-art pretrained models.
- *trl (Apache-2.0 license)*, a full stack library that provides a set of tools to train transformer language models with Reinforcement Learning. The library is built on top of the Transformers library.

D Scaling of Translation Model

We also conducted ablation experiments comparing the translation models used during the preference estimation phase. In addition to NLLB-600M-distilled, we employed NLLB-1.3B-distilled and NLLB-3.3B to estimate the preference and repeat the experiments. The results are shown in Figure 5, where we find that all three models achieved impressive results, showing stable improvements over the Base Model and ChatGPT across ten languages. These results demonstrate that our method is robust for translation models of various sizes.

Model	Bn	Th	Sw	Ja	Zh	Ru	De	Es	Fr	En	Avg
GPT-3.5-Turbo	36.2	42.6	47.2	58.1	60.6	42.6	41.5	54.9	39.4	70.6	49.4
MAmmoTH 7B	1.7	4.7	4.5	23.5	29.9	26.9	34.7	37.9	36.5	42	24.2
WizardMath 7B	9.8	13	6.2	31.8	36	34.3	34.1	39.9	36.5	45.4	28.7
MetaMath 7B	13.2	19.2	12.1	50.7	53.3	52.9	54.2	56.5	56.7	63.3	43.2
MathOctopus 7B	26.6	30.9	34.3	40.9	44.4	36.0	32.6	42.0	36.2	46.9	37.1
+ m-RFT	38.0	42.9	41.8	48.2	51.6	45.2	42.9	49.3	42.9	51.2	45.4
+ MAPO-DPO(ours)	41.8	45.8	46.9	52.9	54.4	49.9	50.7	54.0	51.4	55.9	50.4
MetaMathOctopus 7B	24.5	35.2	36.2	44.1	45.8	39.2	32.6	45.2	36.7	52.5	39.2
+ m-RFT	40.3	45.0	45.2	51.4	57.8	51.6	51.6	58.8	50.1	65.3	51.7
+ MAPO-DPO(ours)	40.7	46.0	45.0	58.2	59.3	53.1	51.4	57.4	52.0	66.1	52.9
MistralMathOctopus 7B	45.2	50.7	49.5	56.5	65.3	59.1	51.4	62.1	53.9	74.6	56.8
+ m-RFT	52.7	60.6	61.0	67.2	70.8	65.9	61.2	71.6	64.4	78.3	65.4
+ MAPO-DPO(ours)	**62.3**	**64.6**	**61.6**	**72.1**	**75.1**	**68.0**	**69.3**	**74.4**	**74.2**	**78.5**	**70.0**
MAmmoTH 13B	6.8	10.5	10.4	31.6	38	41.1	41.4	43.3	42.6	46.7	31.2
WizardMath 13B	9.8	14.3	12.4	30.5	39.0	36.5	35.2	43.9	39.2	47.8	30.9
MetaMath 13B	10.9	16.0	16.0	55.2	57.4	56.5	58.9	60.6	58.0	64.2	45.4
MathOctopus 13B	42.4	39.2	44.8	38.8	49.6	45.2	48.4	53.6	43.2	54.8	46.0
+ m-RFT	44.1	49.9	51.0	51.0	55.6	49.0	47.1	53.3	46.1	57.4	50.5
+ MAPO-DPO(ours)	49.0	53.3	52.2	55.0	57.4	53.3	52.2	55.9	50.7	58.8	53.8
MetaMathOctopus 13B	34.4	42.8	41.6	49.2	52.8	54.4	54.4	59.2	53.6	71.6	51.4
+ m-RFT	40.1	51.6	47.5	60.5	62.0	58.9	54.8	62.5	54.2	66.5	55.9
+ MAPO-DPO(ours)	**52.9**	**55.4**	**55.0**	**67.2**	**65.0**	**54.8**	**54.4**	**65.2**	**57.4**	**70.4**	**59.8**

Table 8: Model Performances on MNumGLUESub test set. “Avg” represents the average performance in ten languages and bold text denotes the best results within the same model size.

Model	Bn	Th	Sw	Ja	Zh	Ru	De	Es	Fr	En	Avg
						Benchmark: MSVAMP					
MathOctopus 7B	27.7	35.9	39.4	41.6	42.7	44.2	44.0	45.1	45.3	46.4	41.2
+ PPO LoRA	38.9	47.5	47.1	51.0	51.7	51.1	50.3	51.6	51.4	52.7	49.3
+ DPO Iter1	42.0	53.2	52.7	54.7	56.4	56.9	55.7	58.5	59.3	59.6	54.9
+ DPO Iter2	45.3	53.9	53.8	56.8	58.1	56.6	58.7	59.1	58.5	60.0	56.1
+ DPO Iter3	48.8	55.2	56.0	60.3	58.8	58.3	58.1	59.7	60.8	58.4	57.4
MetaMathOctopus 7B	36.1	47.5	49.4	51.3	54.5	53.6	56.6	60.0	57.2	64.2	53.0
+ PPO LoRA	45.7	52.0	52.2	61.2	58.3	57.0	58.7	60.7	61.7	67.5	57.5
+ DPO Iter1	44.1	58.2	59.0	60.3	62.3	63.5	65.1	63.7	64.2	70.2	61.1
+ DPO Iter2	48.5	61.8	59.2	64.3	64.3	64.4	65.0	66.1	65.4	70.9	63.0
+ DPO Iter3	50.1	61.6	61.7	65.9	65.7	64.8	68.4	68.5	68.6	71.6	64.7
						Benchmark: MGSM					
MathOctopus 7B	29.2	33.6	36.4	35.2	39.2	38.8	44.8	42.4	43.2	52.0	39.5
+ PPO LoRA	31.2	38.4	38.4	37.2	43.6	35.2	46.0	44.0	38.8	51.2	40.4
+ DPO Iter1	29.2	36.4	35.6	35.6	41.6	38.4	40.8	42.0	37.6	46.8	38.4
+ DPO Iter2	30.4	36.0	37.6	38.0	45.2	39.6	42.0	47.6	41.2	45.2	40.3
+ DPO Iter3	30.8	38.0	37.6	45.2	47.2	42.0	45.2	43.2	40.8	45.6	41.6
MetaMathOctopus 7B	25.6	42.8	36.4	40.0	46.4	46.8	49.6	54.4	46.4	66.4	45.5
+ PPO LoRA	36.0	41.2	41.6	46.4	54.8	53.6	54.0	55.6	51.6	68.0	50.3
+ DPO Iter1	32.8	43.2	40.4	48.8	49.2	52.8	54.4	52.8	50.0	64.8	48.9
+ DPO Iter2	34.0	48.0	45.2	40.4	54.0	52.0	50.8	54.0	49.2	70.4	49.8
+ DPO Iter3	36.0	44.8	44.8	47.6	55.2	53.6	53.6	56.8	52.4	70.8	51.6
						Benchmark: MNumGLUESub					
MathOctopus 7B	26.6	30.9	34.3	40.9	44.4	36.0	32.6	42.0	36.2	46.9	37.1
+ PPO LoRA	34.3	41.1	40.7	45.6	49.3	39.5	34.3	46.7	36.3	51.6	41.9
+ DPO Iter1	32.6	41.4	42.0	42.7	45.0	42.2	40.7	47.8	40.9	47.8	42.3
+ DPO Iter2	35.6	39.7	42.2	45.6	50.7	44.1	42.6	49.3	42.9	49.7	44.2
+ DPO Iter3	41.8	45.8	46.9	52.9	54.4	49.9	50.7	54.0	51.4	55.9	50.4
MetaMathOctopus 7B	34.7	41.4	37.9	47.8	54.2	45	43.1	52.5	45.0	60.8	46.3
+ PPO LoRA	42.4	46.7	45.4	55.2	58.9	45.0	41.6	55.7	44.6	64.6	50.0
+ DPO Iter1	37.7	46.3	43.5	54.8	58.4	50.5	52.2	60.1	51.6	62.1	51.7
+ DPO Iter2	40.7	46.9	45.0	53.5	58.4	50.8	51.4	58.4	49.9	64.6	52.0
+ DPO Iter3	40.7	46.0	45.0	58.2	59.3	53.1	51.4	57.4	52.0	66.1	52.9

Table 9: Model Performances on three benchmarks. We report the results of PPO LoRA and each round in Iterative DPO. "Avg" represents the average performance in ten languages