Question Translation Training for Better Multilingual Reasoning

Wenhao Zhu[1], **Shujian Huang**[1], **Fei Yuan**[2], **Shuaijie She**[1], **Jiajun Chen**[1], **Alexandra Birch**[3]
[1] National Key Laboratory for Novel Software Technology, Nanjing University
[2] Shanghai AI Lab [3] School of Informatics, University of Edinburgh
zhuwh@smail.nju.edu.cn, huangsj@nju.edu.cn, yuanfei@pjlab.org.cn, shesj@smail.nju.edu.cn
chenjj@nju.edu.cn, a.birch@ed.ac.uk

Abstract

Large language models show compelling performance on reasoning tasks but they tend to perform much worse in languages other than English. This is unsurprising given that their training data largely consists of English text and instructions. A typical solution is to translate instruction data into all languages of interest, and then train on the resulting multilingual data, which is called translate-training. This approach not only incurs high cost, but also results in poorly translated data due to the non-standard formatting of mathematical chain-of-thought. In this paper, we explore the benefits of question alignment, where we train the model to translate reasoning questions into English by finetuning on X-English parallel question data. In this way we perform targeted, in-domain language alignment which makes best use of English instruction data to unlock the LLMs' multilingual reasoning abilities. Experimental results on LLaMA2-13B show that question alignment leads to consistent improvements over the translate-training approach: an average improvement of 11.3% and 16.1% accuracy across ten languages on the MGSM and MSVAMP multilingual reasoning benchmarks[1].

1 Introduction

Large language models have recently shown a strong ability to reason in English, but performance in other languages, especially more distant languages, still trails far behind (Shi et al., 2022; Huang et al., 2023). It is unsurprising, considering that their training data is predominantly composed of English text and instructions (Blevins and Zettlemoyer, 2022; Touvron et al., 2023; Wang et al., 2023). To elicit LLM's multilingual performance, previous approach typically follows the translate-training paradigm (Chen et al., 2023), which first translates English instruction data into non-English with a translation engine and then uses the multilingual data for instruction-tuning.

However, the translate-training has the following drawbacks: (1) translating English training data to numerous non-English languages incurs significant translation cost, especially considering the constant addition of large and complex instruction tuning sets (Yuan et al., 2023; Yu et al., 2023). (2) Additionaly, it is hard for the translation engine to accurately translate lengthy, logical texts containing mathematical symbols in chain-of-thought (CoT) responses, which can compromise the quality of translated data (evidence are shown in Appendix A). Consequently, we explore the following research question in this paper: *Can we unlock the LLM's multilingual reasoning ability by teaching it to translate reasoning questions into English?*

In this paper, we focus on the multilingual mathematical reasoning task and explore the benefits of question alignment (QAlign), where we fine-tune the pre-trained LLM to translate reasoning questions into English with X-English parallel question data. This targeted, in-domain language alignment enables the subsequent effective utilization of English instruction data to unlock LLMs' multilingual reasoning abilities. Following question alignment, we implement response alignment by further fine-tuning the language-aligned LLM with cutting-edge English instruction data. Even when only English supervised data is available, our alignment-enhanced LLM can achieve superior performance on non-English tasks with its transferable English expertise.

To demonstrate the advantages of question alignment, we conduct experiments on challenging multilingual mathematical reasoning benchmarks, MGSM (Shi et al., 2022) and MSVAMP (Chen et al., 2023). We use two of the most advanced open-source LLMs, LLaMA2-7B and LLaMA2-13B (Touvron et al., 2023), as base models. Exper-

[1] The project will be available at: https://github.com/NJUNLP/QAlign.

Findings of the Association for Computational Linguistics: ACL 2024, pages 8411–8423
August 11-16, 2024 ©2024 Association for Computational Linguistics

iment results show that the inclusion of the question alignment stage brings an average improvement of up to 13.2% in multilingual performance. The performance improvement on low-recourse languages, e.g. Thai and Swahili, can be 30%-40%. Compared to the translate-training baseline, MathOctopus (Chen et al., 2023), which tuned with a multilingual version of GSM8K dataset , our alignment-enhanced LLMs achieves average performance improvement of 9.6% (7B) and 11.3% (13B) on MGSM. On the out-of-domain test set MSVAMP, our fine-tuned LLMs achieve 13.1% (7B) and 16.1% (13B) average accuracy improvement, also demonstrating our approach is robust to domain shift. In general, we observe that although incorporating translated instruction data does benefit multilingual performance, our question alignment strategy provides a more efficient and effective choice. In our analysis, we also present the effects of other implementations for performing language alignment and illustrate the importance of choosing the appropriate translation direction and domain during this phase of training.

The main contributions of this paper can be summarized as:

- We present a novel X-English question alignment finetuning step which performs targeted language alignment for best use of the LLMs English reasoning abilities.
- We fine-tune open-source LLMs, LLaMA2-7B/13B, into strong multilingual reasoners, which beat the translate-training baseline by 9.6% (7B) and 11.3% (13B) on MGSM, by 13.1% (7B) and 16.1% (13B) on MSVAMP.
- We explore language alignment with other language directions (English-X), types and domains of data, and confirm our intuition that in fact X-English questions perform best.

2 Related Work

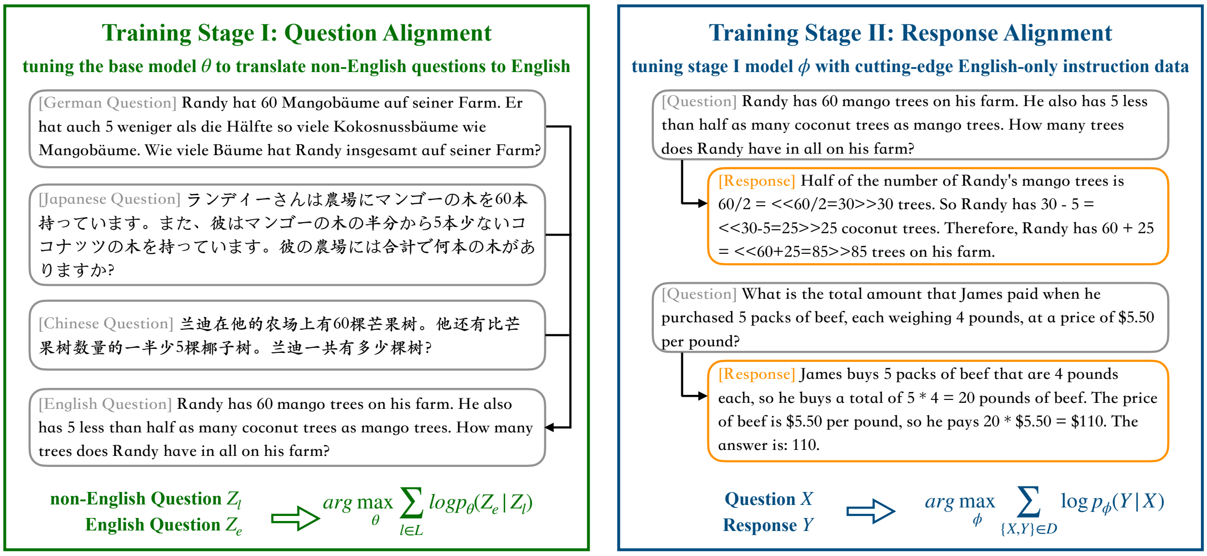
Large language model With a large number of parameters pre-trained on a large-scale corpora, large language models can memorize vast amounts of knowledge (Roberts et al., 2020) and acquire emergent abilitie, such as in-context learning (Brown et al., 2020), CoT generation (Wei et al., 2022b). Then, to better align the behavior of LLMs with human expectations, Wei et al. (2022a) propose instruction-tuning, training LLM to generate desired response based on the given instruction. Subsequently, many efforts are put into creating effective instruction data to further unlock LLM's potential (Wang et al., 2022; Taori et al., 2023; Longpre et al., 2023; Wang et al., 2023). However, since the proposed instruction datasets consist mainly of English, the directly fine-tuned LLMs struggle on non-English languages, especially on those languages that are dissimilar to English (Huang et al., 2023; Zhu et al., 2023; Chen et al., 2023).

Multilingual mathematical reasoning Mathematical reasoning is a challenging and representative task for evaluating the intelligence of LLMs (Ahn et al., 2024), where LLMs need to understand the given math question and produce a numerical answer through step-by-step reasoning. Shi et al. (2022) expanded the scope to a multilingual context by translating English math questions from the GSM8K test set (Cobbe et al., 2021) into non-English languages, thereby creating a multilingual benchmark called MGSM.

Subsequently, many efforts are put into enhancing LLM's multilingual reasoning capabilities, which can be categorized into two approaches: prompting close-source LLMs and instruction-tuning open-source LLMs. In the first approach, Qin et al. (2023) and Huang et al. (2023) carefully craft prompts for close-source LLMs like ChatGPT (OpenAI, 2022). Their strategy involves first prompting the LLM to explicitly translate non-English questions into English, then ask the model to solve the translated problem instead. However, the effectiveness of these prompting methods are not well-examined on open-source LLMs. And it remains an open challenge to equip open-sourced LLMs with strong multilingual mathematical problem-solving skills.

In the second approach, Chen et al. (2023) follow the translate-training method (Artetxe et al., 2023). Initially, they translate English instruction data in GSM8K into non-English with ChatGPT, followed by employing multilingual data for instruction-tuning. Moreover, Chen et al. (2023) investigate cross-lingual training strategies such as mixing questions and CoT responses in different languages, but fail to achieve consistent improvement. Although the translate-training approach is effective, it incurs high translation cost and is error-prone[2]. It also becomes increasingly impractical

[2] We analyze the errors in the translated dataset from Chen et al. (2023) and present both quantitative and qualitative results in Appendix A.


Training Stage I: Question Alignment
tuning the base model θ to translate non-English questions to English
non-English Question Z_l
English Question Z_e
$arg\max_{\theta} \sum_{l \in L} logp_{\theta}(Z_e|Z_l)$
Training Stage II: Response Alignment
tuning stage I model ϕ with cutting-edge English-only instruction data
Question X
Response Y
$arg\max_{\phi} \sum_{\{X,Y\} \in D} \log p_{\phi}(Y|X)$

[German Question] Randy hat 60 Mangobäume auf seiner Farm. Er hat auch 5 weniger als die Hälfte so viele Kokosnussbäume wie Mangobäume. Wie viele Bäume hat Randy insgesamt auf seiner Farm?

[Japanese Question] ランディーさんは農場にマンゴーの木を60本持っています。また、彼はマンゴーの木の半分から5本少ないココナッツの木を持っています。彼の農場には合計で何本の木がありますか?

[Chinese Question] 兰迪在他的农场上有60棵芒果树。他还有比芒果树数量的一半少5棵椰子树。兰迪一共有多少棵树?

[English Question] Randy has 60 mango trees on his farm. He also has 5 less than half as many coconut trees as mango trees. How many trees does Randy have in all on his farm?

[Question] Randy has 60 mango trees on his farm. He also has 5 less than half as many coconut trees as mango trees. How many trees does Randy have in all on his farm?

[Response] Half of the number of Randy's mango trees is 60/2 = <<60/2=30>>30 trees. So Randy has 30 - 5 = <<30-5=25>>25 coconut trees. Therefore, Randy has 60 + 25 = <<60+25=85>>85 trees on his farm.

[Question] What is the total amount that James paid when he purchased 5 packs of beef, each weighing 4 pounds, at a price of $5.50 per pound?

[Response] James buys 5 packs of beef that are 4 pounds each, so he buys a total of 5 * 4 = 20 pounds of beef. The price of beef is $5.50 per pound, so he pays 20 * $5.50 = $110. The answer is: 110.

Figure 1: Illustration of our devised two-step training framework. At training stage I (question alignment), we use a set of multilingual questions for translation training. At training stage II (response alignment), we use cutting-edge English-only instruction data for fine-tuning. Due to the established language alignment in stage I, we can utilize LLM's expertise in English to enhance its performance on non-English tasks.

to translate vast quantities of augmented data into numerous languages, especially considering recent findings that augmented training data, e.g., METAMATHQA (Yu et al., 2023)—which is 50 times larger than GSM8K—greatly enhances LLM's reasoning skills. Without relying on translated CoT responses, in this paper, we present a novel question alignment technique to utilize cutting-edge English-only supervised data to boost open-source LLM's performance on multilingual reasoning tasks.

3 Methodology

An illustration of our devised method is shown in Figure 1. The key idea of our approach is strengthening language alignment within LLM before exposing it to English instruction-response pairs. By doing so, we can utilize LLM's expertise in English to enhance its performance on non-English tasks. Below we introduce the two training stages of our framework: question alignment (§3.1) and response alignment (§3.2).

3.1 Stage I: Question Alignment

It has been found that directly fine-tuning LLMs with English instruction data does not help to improve their performance on non-English tasks (Chen et al., 2023). We suggest that this issue may arise from the insufficient alignment of multiple languages within the LLM. Ideally, in a well-aligned LLM, proficiency in one language, like English, could easily transfer to other languages.

To improve the alignment of non-English languages with English, we devise a translation task **QAlign**: training LLM on translating questions from non-English into English. Specifically, given a group of multilingual questions, the optimization objective can be written as:

$$\arg\max_{\theta} \sum_{l \in \mathcal{L}} \log p_{\theta}(\mathcal{Z}_e | \mathcal{Z}_l)$$

where θ denotes the parameters of the base model. $\mathcal{Z}_l$ and $\mathcal{Z}_e$ denote non-English and English questions respectively and $\mathcal{L}$ is the set of considered non-English languages. With this training objective, we equip the LLM with an implicit bias to relate non-English questions with their English counterparts when performing non-English tasks.

Note that this stage only relies on multilingual questions rather than translated CoT responses. Basically, acquiring multilingual questions is more feasible than obtaining accurate multilingual CoT responses, because translation engines often struggle to precisely translate lengthy, logical texts containing mathematical symbols (quantitative evidence are shown in Appendix A).

In this translation task, the domain of translation data is also an important factor to consider. In subsequent experiments, we demonstrate that using multilingual questions as translation data is more effective than employing general domain translation corpora.

3.2 Stage II: Response Alignment

After question alignment, we train LLM with specialized instruction-response pairs to unlock its potential on multilingual mathematical reasoning tasks. Specifically, we consider two data scenarios: monolingual supervision setting and mixed supervision setting.

Monolingual supervision setting In this setting, we employ English-only instruction data for response alignment, because the cutting-edge instruction datasets are often available only in English. During training, we follow the standard implementation (Wei et al., 2022a) and finetune the language-aligned LLM to maximize the generetive probability of the response $\mathcal{Y}$ given the question $\mathcal{X}$:

$$\arg\max_{\phi} \sum_{\{\mathcal{X},\mathcal{Y}\}\in\mathcal{D}} \log p_{\phi}(\mathcal{Y}|\mathcal{X})$$

Where ϕ denotes the parameters of the stage I model and $\mathcal{D}$ denotes the instruction dataset. Although the training only utilizes English supervision, the previously established language alignment allows the LLM's English proficiency to be shared across multiple languages.

Mixed supervision setting While our framework is primarily designed for the scenario where only English instruction data is available, it can also leverage additional multilingual supervision, when available, to achieve even higher multilingual performance. For instance, this multilingual dataset could be a translated version of a subset of large-scale English data. In this scenario, given a set of additional multilingual superivsed data $\mathcal{M}$, we sequentially fine-tune the stage I model on $\mathcal{M}$ and then on the English instruction data $\mathcal{D}$. Subsequent experiment results show that this training recipe can further improve the LLM's multilingual reasoning capabilities.

4 Experiment Setting

Base LLM In our experiments, we use two of the most advanced open-source LLMs, LLaMA2-7B and LLaMA2-13B as the base model.

Training Dataset In the question alignment stage, we utilize multilingual questions from GSM8KINSTRUCT[3] (Chen et al., 2023). During the response alignment stage, we employ the cutting-edge English-only dataset METAMATHQA as monolingual supervision, which is built upon English dataset GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) by performing data augmentation, such as rephrasing questions and enriching answers. In the mixed supervision setting, we employ both METAMATHQA and GSM8KINSTRUCT. Dataset statistics are reported in Table 1.

Dataset	Usage	# Lang	# Sample
METAMATHQA	Training	1	395,000
GSM8KINSTRUCT	Training	10	73,559
MGSM	Evaluation	10	2,500
MSVAMP	Evaluation	10	10,000

Table 1: Statistics of involved datasets. "# Lang" denotes the number of languages covered by the dataset and "# Sample" refers to the total number of samples it contains.

Training Details We use *stanford_alpaca*[4] as our code base. We use consistent training hyperparameters across two stages of training. At each stage, we fine-tune LLM's full parameters for 3 epoch on eight NVIDIA A100 GPUs. The learning rate is set to 2e-5, with a batch size of 128.

Baseline Systems For comparison, we consider following systems which are instruction-tuned from LLaMA2 with diverse training recipes:

- **SFT** (Touvron et al., 2023), which is instruction-tuned with basic GSM8K.
- **RFT** (Yuan et al., 2023), which is instruction-tuned with an augmented GSM8K training dataset, using rejection sampling techniques.
- **MAmmoTH** (Yue et al., 2023), which is instruction-tuned with GSM8K and a collection of math instruction datasets.
- **WizardMath** (Luo et al., 2023), which is constructed using reinforcement learning on GSM8K and MATH.
- **MathOctopus** (Chen et al., 2023), which is instruction-tuned with a multilingual version of GSM8K dataset, representing a standard implementation of translate-training approach. We also reproduce this model in our experiments, denoted as **MultiReason**.

[3] GSM8KINSTRUCT is a multilingual dataset that extends the English instruction dataset GSM8K by translating English instructions and CoT responses into nine non-English languages with ChatGPT.

[4] https://github.com/tatsu-lab/stanford_alpaca

System (7B)	Bn	Th	Sw	Ja	Zh	De	Fr	Ru	Es	En	Avg.
SFT† (Touvron et al., 2023)	3.2	4.8	5.2	15.2	22.4	37.2	34.4	28.0	32.4	43.2	22.6
RFT† (Yuan et al., 2023)	2.4	2.0	2.8	6.8	16.8	33.6	34.0	29.2	34.0	44.8	20.6
MAmmoTH† (Yue et al., 2023)	3.6	4.8	2.4	10.8	17.2	33.2	32.8	26.0	32.4	49.6	21.3
WizardMath† (Luo et al., 2023)	2.0	4.0	3.4	24.0	22.4	30.4	30.4	30.8	34.8	47.6	23.0
MathOctopus† (Chen et al., 2023)	28.8	34.4	39.2	36.0	38.4	44.8	43.6	39.6	42.4	52.4	40.0
MetaMath (Yu et al., 2023)	6.4	4.0	3.2	39.2	38.8	56.8	52.8	47.2	58.0	63.2	37.0
MultiReason	26.8	36.0	36.8	33.2	42.4	42.8	40.8	42.4	42.8	47.2	39.1
MonoReason	7.6	5.6	5.2	34.0	45.2	54.0	**56.8**	51.6	58.8	65.5	38.4
QAlign→MonoReason (Ours)	**32.4**	**39.6**	**40.4**	**44.0**	**48.4**	**54.8**	**56.8**	**52.4**	**59.6**	**68.0**	**49.6**

System (13B)	Bn	Th	Sw	Ja	Zh	De	Fr	Ru	Es	En	Avg.
SFT† (Touvron et al., 2023)	6.0	6.8	7.6	25.2	32.8	42.8	40.8	39.2	45.2	50.4	29.7
RFT† (Yuan et al., 2023)	3.2	4.4	3.6	26.4	33.6	38.4	44.8	41.6	46.8	52.0	29.5
MAmmoTH† (Yue et al., 2023)	3.6	5.2	1.6	19.2	31.2	45.6	39.6	36.8	50.0	56.4	28.9
WizardMath† (Luo et al., 2023)	6.4	5.6	5.6	22.0	28.0	40.4	42.0	34.4	45.6	52.8	28.3
MathOctopus† (Chen et al., 2023)	35.2	46.8	42.8	43.2	48.8	44.4	48.4	47.6	48.0	53.2	45.8
MetaMath (Yu et al., 2023)	11.6	6.4	7.6	42.8	49.2	**64.8**	**65.2**	63.6	65.2	67.2	44.4
MultiReason	37.6	42.2	44.0	43.2	53.6	47.6	54.0	48.0	54.8	56.4	48.1
MonoReason	12.4	11.2	6.4	42.0	46.0	64.0	62.4	61.6	64.8	68.4	43.9
QAlign→MonoReason (Ours)	**38.4**	**49.6**	**46.0**	**52.4**	**59.2**	62.0	62.4	**64.4**	**67.2**	**69.2**	**57.1**

Table 2: Results on MGSM dataset. "Avg." represents the average multilingual performance and bold text denotes the highest score among systems of the same size. The dagger symbol denotes that the results for these models are taken from the published results of Chen et al. (2023).

- **MetaMath**, which is instruction-tuned with METAMATHQA (Yu et al., 2023). It is currently the most powerful English instruction data for mathematical reasoning. We also reproduce this model in our experiments, denoted as **MonoReason**.

Among these baseline systems, most models are tuned with English data and only MathOctopus and MultiReason are tuned with multilingual data.

Evaluation Dataset To assess LLMs' performance on multilingual mathematical reasoning[5], we employ the benchmark dataset MGSM (Shi et al., 2022). We also evaluate the robustness of LLMs using an out-of-domain test set MSVAMP (Chen et al., 2023). In our experiments, we report LLM's answer accuracy in a zero-shot and greedy decoding setting. Specifically, we use evaluation scripts[6] provided by Chen et al. (2023) and measure answer accuracy by comparing the last numerical number that appears in the LLM-generated response with the gold answer.

5 Main Results

In this section, we report our experiment results and introduce our main findings.

5.1 Monolingual Supervision Setting

Question alignment stage enables LLM's proficiency in English to be transferred to non-English tasks. Experiment results on the MGSM dataset are presented in Table 2. We can see that LLMs trained with augmented English data (RFT, MAmmoTH, WizardMath, MetaMath and MonoReason) typically underperform on non-English tasks, despite showing improved performance in English compared to SFT model. The multilingual MathOctopus outperforms existing open-source models in terms of multilingual performance. However, as we have discussed, the translated dataset can be out-dated quickly and keeping translating cutting-edge English instuction can also be prohibitive due to the high translation cost.

Unlike the translate-training approach, our framework can easily utilize the most advanced English instruction data, e.g., METAMATHQA. With the question alignment stage (QAlign), we successfully transfer model's proficiency in English to non-English languages. On average, this leads to a 11.2% increase in accuracy for the 7B model and a 13.2% increase in accuracy for the

[5] In this paper, we evaluate LLMs' reasoning performance on ten languages: Bengali (Bn), Thai (Th), Swahili (Sw), Japanese (Ja), Chinese (Zh), German (De), French (Fr), Russian (Ru), Spanish (Es) and English (En).

[6] `https://github.com/microsoft/MathOctopus`

System (7B)	Bn	Th	Sw	Ja	Zh	De	Fr	Ru	Es	En	Avg.
SFT† (Touvron et al., 2023)	11.5	18.2	17.2	31.6	35.2	39.0	39.1	39.1	39.2	38.8	30.9
RFT† (Yuan et al., 2023)	7.7	16.9	14.9	33.9	34.9	40.8	41.5	39.5	42.5	42.7	31.3
MAmmoTH† (Yue et al., 2023)	4.3	6.3	4.2	26.7	26.8	39.6	39.9	33.7	42.9	45.1	26.3
WizardMath† (Luo et al., 2023)	16.1	17.0	10.3	37.9	36.3	39.2	37.7	37.4	44.8	48.5	32.5
MathOctopus† (Chen et al., 2023)	31.8	39.3	43.4	41.1	42.6	48.4	50.6	46.9	49.4	50.7	44.1
MetaMath (Yu et al., 2023)	14.2	17.8	16.5	53.2	53.1	61.4	60.7	58.9	61.2	65.5	46.3
MultiReason	27.6	36.5	42.4	40.9	43.2	44.3	46.7	42.3	45.5	48.0	41.3
MonoReason	15.0	17.1	15.4	51.9	54.4	60.9	62.2	59.3	**63.3**	**65.5**	46.2
QAlign→MonoReason (Ours)	**41.7**	**47.7**	**54.8**	**58.0**	**55.7**	**62.8**	**63.2**	**61.1**	**63.3**	65.3	**57.2**

System (13B)	Bn	Th	Sw	Ja	Zh	De	Fr	Ru	Es	En	Avg.
SFT† (Touvron et al., 2023)	13.9	23.4	19.8	41.8	43.3	46.2	47.8	47.8	46.1	50.9	38.1
RFT† (Yuan et al., 2023)	12.2	24.8	19.4	42.4	42.3	45.1	45.2	46.5	45.6	47.1	37.1
MAmmoTH† (Yue et al., 2023)	5.0	13.7	12.9	42.2	47.7	52.3	53.8	50.7	53.9	53.4	38.6
WizardMath† (Luo et al., 2023)	13.7	16.3	12.5	29.5	37.0	48.7	49.4	43.8	49.4	56.3	35.7
MathOctopus† (Chen et al., 2023)	35.2	41.2	46.8	39.2	52.0	47.2	48.0	45.6	53.2	56.4	46.5
MetaMath (Yu et al., 2023)	14.6	15.7	17.4	57.0	56.6	67.3	64.7	63.7	65.9	67.7	49.1
MultiReason	35.0	41.3	44.6	49.9	48.1	53.3	53.2	51.6	52.5	54.5	48.4
MonoReason	20.6	20.5	19.1	57.0	58.8	68.4	**68.1**	**67.5**	**68.9**	**68.9**	51.8
QAlign→MonoReason (Ours)	**49.2**	**55.5**	**55.2**	**64.3**	**63.8**	**69.5**	**68.1**	66.4	66.4	67.6	**62.6**

Table 3: Results on MSVAMP dataset. "Avg." represents the average multilingual performance and bold text denotes the highest score among systems of the same size. The dagger symbol denotes that the results for these models are taken from the published results of Chen et al. (2023).

13B model. These substantial improvements on non-English languages significantly reduce LLM's performance gap between non-English and English tasks, thereby demonstrating the effectiveness of our devised method.

After question alignment, our fine-tuned LLM surpasses the translate-training baseline by a large margin More importantly, we observe that after question alignment, our fine-tuned LLM surpasses the translate-training baseline (MathOctopus) by a large margin. By transferring the model's expertise in English to non-English scenarios, our approach outperforms MathOctopus by an average margin of +9.6% for the 7B model and +11.3% for the 13B model. These results again demonstrate the superiority of our method[7].

Our fine-tuned LLMs also exhibit better robustness on the out-of-domain test set Apart from evaluating on MGSM, we further assess the robustness of our LLMs on the out-of-domain test set MSVAMP (Table 10). The findings are generally consistent with those from MGSM dataset. Notably, compared to the unaligned counterpart (MonoReason), our model (QAlign→MonoReason) achieves significant improvement in average multilingual performance, with gains of 11.0% for the 7B model and 10.8% for the 13B model. Our method outperforms the translate-training approach (MathOctopus) by an even larger margin here, showing increases of 13.1% for the 7B model and 16.1% for the 13B model, which shows its more generalized and robust performance.

5.2 Mixed Supervision Setting

Incorporating multilingual supervised data into our framework can achieve a higher ceiling for multilingual performance Although our framework does not rely on the multilingual supervised data, we can utilize such data to attain a higher level of multilingual performance if a multilingual dataset is available. In this mixed supervision setting, we first tune the stage I model (7B) with multilingual GSM8KInstruct and then tune it with English data MetaMathQA. The experiment results on MGSM are depicted in Figure 2. We find that incorporating additional multilingual supervision further leads to an average performance gain of 2.1% on multilingual tasks. Compared to the data mixing baseline (MonoReason+MultiReason), our

[7] In Appendix B, we also report the results of using question translation data for stage I training and multilingual instruction data for stage II training. This provides a more direct comparison (QAlign→MultiReason vs. MultiReason), and the added question alignment stage also improves multilingual performance in this setting.

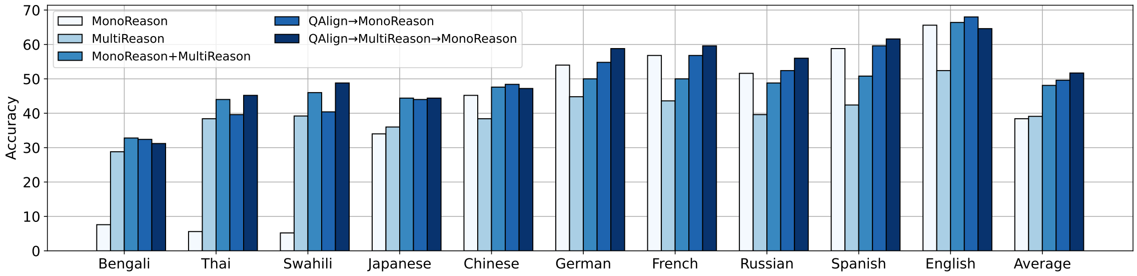


Figure 2: Effects of tuning language-aligned LLM with mixed supervised data. Generally, incoporating multilingual supervised data into our framework can achieve a higher ceiling for average multilingual performance.

Data	Direction	mGSM		mSVAMP	
		Non-En	En	Non-En	En
Question	X→En	47.6	68.0	56.5	65.3
Question	En→X	36.2	68.0	48.3	64.4
Response	X→En	46.4	67.2	52.1	64.9
Response	En→X	42.8	68.0	49.0	63.9
Flores-101	X→En	36.3	68.0	46.8	65.4

Table 4: Effects of using different translation training data for stage I training. "X→En" and "En→X" represents translating from non-English to English and translating English to non-English respectively. "Non-En" denotes LLM's average performance on non-English languages. Among these implementations, training LLM to translate non-English questions to English is the best one.

approach demonstrate an average improvement of 3.6%, with significant advantages in high-resource languages such as Spanish, Russian, German, and French.

6 Analysis

6.1 Ablation study

Impact of using different translation training data During the question alignment stage, we implement the translation task by training LLMs on translating questions from non-English to English. Now we present the ablation study to show the effects of alternative implementations (Table 4). while different implementations yield similar performance in English, their impact on non-English peformance varies significantly. For instance, training LLMs on reverse translation tasks greatly degenerates non-English performance (*Question*:En→X, *Response*:En→X). Training LLM on translating CoT responses from non-English to English (*Response*:X→En) also results in lower performance compared to our original implementation. We suggest that this is because noises in the translated CoT responses compromise the data quality. Training the LLM with translation data from commonly-used corpora, such as FLORES[8], does not work as well, indicating that the domain of the translation data is another crucial factor in establishing language alignment.

Implementation	mGSM		mSVAMP	
	Non-En	En	Non-En	En
our implementation	47.6	68.0	56.5	65.3
↪ *reversing training order*	2.0	2.8	2.0	2.0
↪ *single-stage training*	3.7	68.0	2.6	65.2

Table 5: Effects of reversing training order and performing single-stage multi-task training. Among these implementations, our original implementation, i.e., performing question alignment at first and then perform response alignment, is the best one.

Impact of manipulating training order We also conduct the ablation study to demonstrate the significance of the training sequence within our proposed framework. As shown in Table 5, reversing the order of the two training stages results in the LLM performing poorly in both English and non-English languages. We observed that an LLM fine-tuned in this manner tends to repeat the question in English when presented with questions in various languages.

When we merge the training datasets from both stages and perform a single-stage, multi-task training, there is a significant drop in non-English performance as well. Although capable of responding to questions in English, the fine-tuned LLM is prone to translating the given non-English questions rather than answering them. These analysis results demonstrate that our design of two-step training framework is non-trivial.

[8] In this ablation study, we take the translation data in the development and test set of FLORES-101 dataset (Goyal et al., 2022) for fine-tuning.

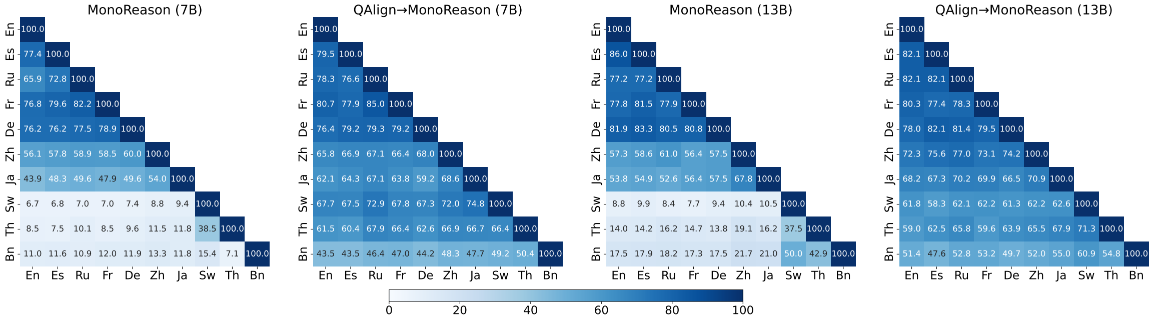


Figure 3: Comparing the prediction consistency of different systems. Darker blue denotes higher level of prediction consistency. Question alignment stage always brings improvement to the consistency of predicted answers.

Method	MGSM		MSVAMP	
	Non-En	En	Non-En	En
MonoReason (7B)				
Direct Inference	35.4	65.5	47.6	68.9
Translate-test	30.8	-	42.3	-
QAlign→MonoReason (7B)				
Direct Inference	47.6	68.0	56.5	65.3
Translate-test	46.6	-	56.6	-

Table 6: Comparison between direct inference and translate-test inference.

Supervision	QAlign	MGSM		MSVAMP	
		Non-En	En	Non-En	En
GSM8K	✗	18.8	43.6	33.6	47.2
GSM8K	✓	26.3	41.6	36.8	47.0
METAMATHQA	✗	35.4	65.6	44.4	65.3
METAMATHQA	✓	47.6	68.0	56.5	65.3

Table 7: Effects of tuning the stage I model (7B) with different English instruction data.

6.2 Prediction Consistency

Another advantage of establishing question alignment is the improvement it brings to the consistency[9] of predicted answers against multilingual queries. This means a higher degree of agreement in answers to the same question posed in different languages. Figure 3 displays the quantified results. In contrast to their unaligned counterparts (MonoReason), our alignment-enhanced LLM (QAlign→MonoReason) usually demonstrate higher answer consistency. This improvement is particularly notable for distant languages, such as Bengali, Thai, Swahili, Japanese, and Chinese. This results can serve as another strong evidence of our successful transfer of LLM's proficiency in English to non-English languages. Appendix C presents some cases to further illustrate the advantages of achieving higher multilingual consistency.

6.3 Question Alignment vs Translate-Test

In our training framework, we implicitly endow the LLM with a bias that associates non-English questions with their English equivalents, sharing similar philosophy with translate-test prompting approach. Thus we discuss the difference between these two approaches here. Experiment results are reported in Table 6. For the MonoReason model, the translate-test approach does not yield any improvement, suggesting that this approach may not be universally applicable solution for open-source LLMs. For our alignment-enhanced QAlign→MonoReason model, direct inference and translate-test prompting achieves similar performance. But considering our approach does not rely on explicitly translating the questions during inference, it will have a more efficient inference process.

6.4 Effects of tuning LLM with different English instruction data

To demonstrate the universal effectiveness of question alignment, we also employ English GSM8K dataset as monolingual supervison and show the results in Table 7. Under different English instruction data, the incorporation of a question alignment stage always boost LLM's non-English performance. These results also highlight the importance of using advanced English instruction data, because achieving better performance in English usually means an improved non-English performance with the help of inner language alignment.

7 Conclusion

In this paper, we introduce a novel question alignment method to empower LLMs on multilingual

[9] Supposing the set of correct predictions in two languages is U and V respectively, we compute the consistency score as $\frac{|U \cap V|}{|U|}$.

mathematical reasoning tasks without requiring multilingual instruction data. Experiment results on mGSM and mSVAMP benchmarks show that our proposed question alignment stage brings an average improvement of up to 13.2% in multilingual performance. Our alignment-enhanced LLM outperforms the unaligned counterpart and the translate-training baseline by a large margin and shows a more robust performance. Generally, our devised method successfully narrows the gap between LLM's performance between English and non-English, showing a new possibility to unlock LLM's capabilities to solve multilingual tasks.

Limitation

Below we discuss potential limitations of our work:

- Chain-of-Thought in English: When receiving non-English questions, our language-aligned LLM typically produces an English CoT before giving the final numerical answer. While the language used for the CoT is not explicitly specified as a requirement for the multilingual mathematical reasoning task, providing a CoT consistent with the query's language could enhance the model's utility.
- Scale of the pre-trained LLM: Our experiment is constrained by available computational resources, leading us to utilize the LLaMA2-7B and LLaMA2-13B models. Should resources allow in the future, we aim to broaden our research to include larger-scale models, such as LLaMA2-70B.

Acknowledgement

We would like to thank the anonymous reviewers for their insightful comments. Shujian Huang is the corresponding author. This work is supported by National Science Foundation of China (No. 62376116, 62176120), the Liaoning Provincial Research Foundation for Basic Research (No. 2022-KF-26-02), research project of Nanjing University-China Mobile Joint Institute. This project has also received funding from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (UTTER grant numbers 10039436). Wenhao Zhu is also supported by China Scholarship Council (No.202306190172).

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. *arXiv preprint arXiv:2305.14240*.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*.

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, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics (TACL)*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*.

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. *arXiv preprint arXiv:2308.09583*.

OpenAI. 2022. https://openai.com/blog/chatgpt.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *arXiv preprint arXiv:2310.14799*.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Conference on Machine Translation (WMT)*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. In *International Conference on Learning Representations (ICLR)*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Supernaturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations (ICLR)*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

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. *arXiv preprint arXiv: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. *arXiv preprint arXiv:2309.05653*.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

A Analyzing the Quality of the Translated Dataset

In the work of (Chen et al., 2023), the authors employ ChatGPT to translate GSM8K into several non-English languages, resulting in the creation of the multilingual dataset GSM8KInstruct. Below we analyze the translation quality of this dataset and highlight the challenges associated with translating complex CoT responses. We evaluate the translation quality of both questions and responses in a reference-free condition with COMETKiwi[10] (Rei et al., 2022). The evaluation results in Table 8 show that the quality of the translated responses is significantly inferior to that of the translated questions. This gap demonstrates the difficulties inherent in translating CoT content.

Table 9 provides some examples of typical translation errors. Based on this analysis, we suggest that constructing a multilingual CoT dataset through a translation engine is fraught with errors and cannot ensure the quality of the dataset. In constrast, our devised framework provides a more effective and efficient solution, which does not require translated multilingual CoT.

B Experiment Results of Using Multilingual Instruction Data for Response Alignment

To more comprehensively illustrate the benefit of the question alignment approach, we use question translation data for stage I training and multilingual instruction data GSM8KInstruct for stage II training (QAlign→MultiReason). We can see that the added question alignment stage also brings improvement on multilingual performance in this setting.

C Benefits of Achieving Higher Multilingual Consistency

Table 11 presents some examples to illustrate the advantages of achieving higher multilingual consistency. In this case, both MonoReason model and QAlign→MonoReason model predict correctly on the English question. However, when presented with the same question expressed in Chinese, our LLM generates a similar reasoning path and still give the right answer, but the unaligned model fails.

D Used Scientific Artifacts

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

- *Stanford Alpaca (Apache-2.0 license)*, a project that aims to build and share an instruction-following LLaMA model.
- *Transformers (Apache-2.0 license)*, a framework that provides thousands of pretrained models to perform tasks on different modalities such as text, vision, and audio.

[10] Specifically, we employ *wmt22-cometkiwi-da* as the evaluation model: https://huggingface.co/Unbabel/wmt22-cometkiwi-da.

Analyzed Data	COMETKiwi (En-X)								
	Bn	Th	Sw	Ja	Zh	De	Fr	Ru	Es
Question Translation	82.22	79.61	82.60	86.64	82.95	83.56	82.29	84.53	85.59
Response Translation	79.92	76.99	76.84	83.34	79.30	78.34	79.98	79.85	79.83

Table 8: Evaluation results of the translation quality of GSM8KINSTRUCT dataset.

Example I: missing certain reasoning step in the translated response

[**English CoT Response**]
Half of the wallet's price is $99 / 2 = $50.
Betty's grandparents gave her $15 * 2 = $30.
In total, Betty has $50 + $15 + $30 = $95.
So she still needs $100 – $95 = $5 which her parents plan to give her for her birthday.

[**Translated Chinese CoT Response**]
钱包价格的一半是100美元/ 2 = 50美元。
(Half the wallet price is $100/2=$50.)
贝蒂的祖父母给了她15美元* 2 = 30美元。
(Betty's grandparents gave her $15*2=$30.)
这意味着，贝蒂还需要50美元- 30美元- 15美元= 5美元。
(This means that Betty still needs $50 - $30 - $15 = $5.)

Example II: distinct reasoning paths exist between responses in English and the translated version

[**English CoT Response**]
The total time he takes to walk to work 3 times a week is 2*3 = 6 hours
The total time he takes to walk to and from work 3 times a week is 6+6 = 12 hours
The total time he takes to bike to work 2 times a week is 2*1 = 2 hours
The total time he takes to bike to and from work 2 times a week is 2+2 = 4 hours
The total time he takes to go to and from work in a week is 12+4 = 16 hours

[**Translated Chinese CoT Response**]
Roque每周走路去上班需要2*3=6小时。
(It takes Roque 2*3=6 hours to walk to work every week.)
Roque每周骑自行车去上班需要1*2=2小时。
(It takes Roque 1*2=2 hours to ride a bicycle to work every week.)
他每周去上班的总时间是6+2=8小时。
(The total time he goes to work every week is 6+2=8 hours.)
他每周回家的总时间也是8小时。
(The total time he spends home every week is also 8 hours.)
因此，Roque每周去上班和回家总共需要8+8=16小时。
(Therefore, it takes Roque a total of 8+8=16 hours per week to get to work and back home.)

Table 9: Examples of failing to accurately translating English CoT response. For illustration, the gray text is the English meaning of the corresponding Chinese sentence.

System (7B)	**Bn**	**Th**	**Sw**	**Ja**	**Zh**	**De**	**Fr**	**Ru**	**Es**	**En**	**Avg.**
MultiReason	26.8	36.0	36.8	33.2	42.4	42.8	**40.8**	**42.4**	42.8	47.2	39.1
QAlign→MultiReason	**31.6**	**36.4**	**38.8**	**38.0**	**43.6**	**45.2**	**40.8**	38.4	**46.8**	**49.2**	**40.9**

Table 10: Results on MGSM dataset. "Avg." represents the average multilingual performance and bold text denotes the highest score among systems of the same size.

Question (English)	**Gold Answer**
There are 27 unicorns left in the world. One third of them are in the Scottish Highlands. Two thirds of the Scottish unicorns are female. How many female Scottish unicorns are there?	6
Chain-of-Thought Response	**Prediction**
[**MonoReason**] One third of the unicorns are in the Scottish Highlands, so there are 27/3 = 9 unicorns in the Highlands. Two thirds of the Scottish unicorns are female, so there are 9 * 2/3 = 6 female Scottish unicorns. The answer is: 6	6 ✓
[**QAlign→MonoReason**] One third of the unicorns are in the Scottish Highlands, so there are 27/3 = 9 unicorns in the Highlands. Two thirds of the unicorns in the Highlands are female, so there are 9 * 2/3 = 6 female unicorns in the Highlands. Therefore, there are 6 female Scottish unicorns. The answer is: 6	6 ✓
Question (Chinese)	**Gold Answer**
世界上剩下27只独角兽，其中三分之一在苏格兰高地。 苏格兰高地的独角兽的三分之二是雌性。苏格兰雌性独角兽有多少只?	6
Chain-of-Thought Response	**Prediction**
[**MonoReason**] If three-fourths of the remaining 27 one-horned animals are in the Scottish Highlands, then there are 27 * 3/4 = 21 one-horned animals in the Scottish Highlands. If two-thirds of the one-horned animals in the Scottish Highlands are female, then there are 21 * 2/3 = 14 female one-horned animals in the Scottish Highlands. Therefore, there are 14 female one-horned animals in the Scottish Highlands. The answer is: 14	14 ✗
[**QAlign→MonoReason**] If there are 27 remaining unicorns in the world, and one-third of them are in the Scottish Highlands, then there are 27/3 = 9 unicorns in the Scottish Highlands. If two-thirds of the unicorns in the Scottish Highlands are female, then there are 9 * 2/3 = 6 female unicorns in the Scottish Highlands. Therefore, there are 6 female unicorns in the Scottish Highlands. The answer is: 6	6 ✓

Table 11: With this case we illustrate the advantages of achieving higher prediction consistency. Given the same question expressed in different languages, our alignment-enhanced model (QAlign→MonoReason) can generate a similar reasoning path and give the right answer consistently.