Cognitive Benefits of Multilingualism in Reasoning Challenges

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have shown impressive performance on reasoning tasks, especially when conducted in English. However, leveraging multilingual capabilities can significantly enhance reasoning effectiveness. In this paper, we comprehensively explore the benefits of multilingualism in reasoning using BenchMAX and highlight a range of intriguing phenomena. Our findings indicate that employing multiple languages can provide additional advantages, with a notably high upper bound for these benefits. This upper bound demonstrates remarkable tolerance for variations in translation quality and language choice, yet it remains sensitive to the methods used for answer selection. Unfortunately, common answer selection strategies often fail to unlock the full potential of multilingualism. Further analysis of the benefits and challenges shows that key languages like Korean and French can enhance the reasoning abilities of various models, and common answer selection struggles because it depends on language combinations and its performance does not improve with more languages. These insights may pave the way for future research aimed at fully harnessing the potential of multilingual reasoning in LLMs [1].

1 Introduction

Large Language Models (LLMs; OpenAI et al., 2024; Gemini, 2024; Team, 2025) excel in reasoning (Wei et al., 2022; Yao et al., 2023; Team, 2025; Li et al., 2025), and *these models tend to achieve higher performance when tasks are presented in English* (Shi et al., 2023; Huang et al., 2022; Fu et al., 2022; She et al., 2024; Etxaniz et al., 2024).

However, reasoning should not be limited to English - being multilingual can boost thinking effectiveness. This intriguing phenomenon has been substantiated in human education by the Ministry of Education of Mali (Bühmann, 2008), as shown in Figure 1, revealing that bilingual education can lead to more than a 30% improvement in mathematics scores compared to monolingual education.

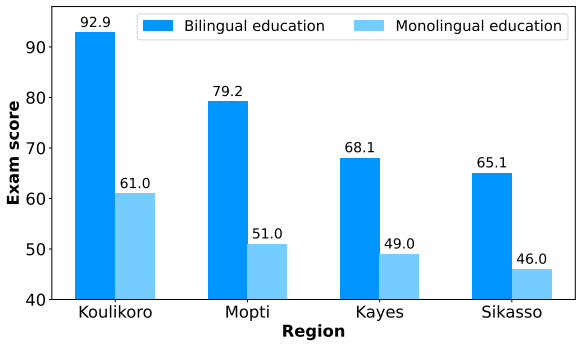

Figure 1: A study from the Ministry of Education of Mali demonstrates that mathematics scores are higher in bilingual schools compared to monolingual schools, highlighting the advantages of multilingualism.

In fact, the training process of LLMs can be seen as a form of multilingual education. These models are built on a robust multilingual foundation (Yuan et al., 2024), developed through extensive exposure to diverse multilingual data and effective vocabulary sharing across various languages. Do existing LLMs recognize that they can, akin to humans, utilize multilingualism to enhance their performance on reasoning-related tasks?

Firstly, we investigate the impact of multilingualism on LLMs reasoning performance by conducting extensive experiments on two reasoning-specific tasks—reasoning-math and reasoning-science—within the BenchMAX framework (Huang et al., 2025). Interestingly, English does not consistently yield the best results in these two tasks, even though these models have been extensively trained on English-centric data.

To further explore the potential benefits of multilingualism on reasoning performance, we conduct a comprehensive comparison of various approaches, including Multilingual, Re-

[1]The code will be made publicly available.

peat, Paraphrase, Repeat-Mix, and Paraphrase-Mix. We maintain strict consistency in other hyper-parameters and evaluate using the Acc@k score (where having one correct answer among k generated answers is considered correct) across multiple experiments. The performance of the Multilingual approach surpasses that of the Repeat, Paraphrase, Repeat-Mix, and Paraphrase strategies. When comparing Repeat-Mix (or Paraphrase-Mix) with Repeat (or Paraphrase), substituting English input with multilingual input yields a significant improvement. These demonstrate that leveraging multilingualism can provide substantial additional benefits in reasoning tasks. Notably, the upper-bound of this advantage exceeds that of Repeat and Paraphrase by as much as 10 Acc@k points.

Surprisingly, this upper-bound is robust with language selection and translation quality. We randomly select four languages, and though none of these languages individually outperforms English, their combination still results in a significant gain. Furthermore, switching from high-quality human translations to machine translation does not lead to any substantial change in performance. However, it is sensitive to answer selection strategy, indicating that selecting the right answer from candidates is essential for unlocking multilingual potential. In our analysis, we propose two possible reasons behind the upper-bound gain, the first being the language-wise correctness correlates with the question difficulty, and the second being existence of key advantaged languages that can compensate for other languages' errors.

We then explore commonly used answer selection methods including Prompt-based-selection and LLM-as-a-judge. Unfortunately, many experimental results reveal that performance gains occur inconsistently across different settings, suggesting that a stable selection method for leveraging multilingualism for enhanced reasoning remains elusive. Our analysis suggests that this is related to the shortcoming of majority voting, and the language bias of prompt-based and LLM-as-a-judge selection.

The main contribution can be summarized as:

- We comprehensively analyze how multilingualism can enhance reasoning capabilities, laying the groundwork for understanding its huge potential.
- We evaluate common answer selection methods and find it is a challenge to tap into the advantage of multilingualism, highlighting the difficulties.
- Extensive experiments reveal the huge gains, point out the limitations of existing methods, and share interesting findings for future research.

2 Related Work

Enhancing LLMs' Reasoning Performance Enhancing reasoning capabilities has emerged as a central challenge in LLM research. Prior work has approached this challenge from three main directions: prompting, pre-training, and post-training methods. In terms of prompting, chain-of-thought (CoT) has proven particularly effective (Wei et al., 2022), enabling models to break down complex problems into intermediate steps and achieve higher reasoning accuracy. For pre-training, recent studies have explored the relationship between code pretraining and reasoning capabilities (Aryabumi et al., 2024), revealing important insights into model behavior. Post-training approaches, including reinforcement learning (Team, 2025) and instruction-tuning (Muennighoff et al., 2025), have also shown promising results in enhancing reasoning performance. Our work complements these studies by focusing on the impact of multilingualism on LLM's reasoning behavior.

Multilingualism in LLM Multilingual capability is crucial in LLM development. Earlier LLMs exhibited unbalanced performance across languages, with non-English CoT reasoning typically underperforming compared to English CoT (Shi et al., 2023; She et al., 2024; Zhu et al., 2024). However, recent advances in pre-trained language models have significantly transformed this landscape. Notably, Huang et al. (2025) demonstrated that state-of-the-art LLMs such as `Qwen` (Qwen Team, 2025) and `LLaMA` (Dubey et al., 2024) achieve superior reasoning accuracy with non-English CoT compared to their English counterparts (Shi et al., 2023). In this paper, we systematically investigate this phenomenon and explore how to leverage multilingual reasoning to probe LLMs' performance ceiling.

3 Multilingualism Empowers Reasoning

3.1 Pilot Study Setup

To examine multilingual reasoning benefits, we analyze LLM responses to questions translated into multiple languages, and ablate the gains of increasing multilingualism versus increasing sampled response numbers (Figure 2). Specifically, we compare following approaches to transform question

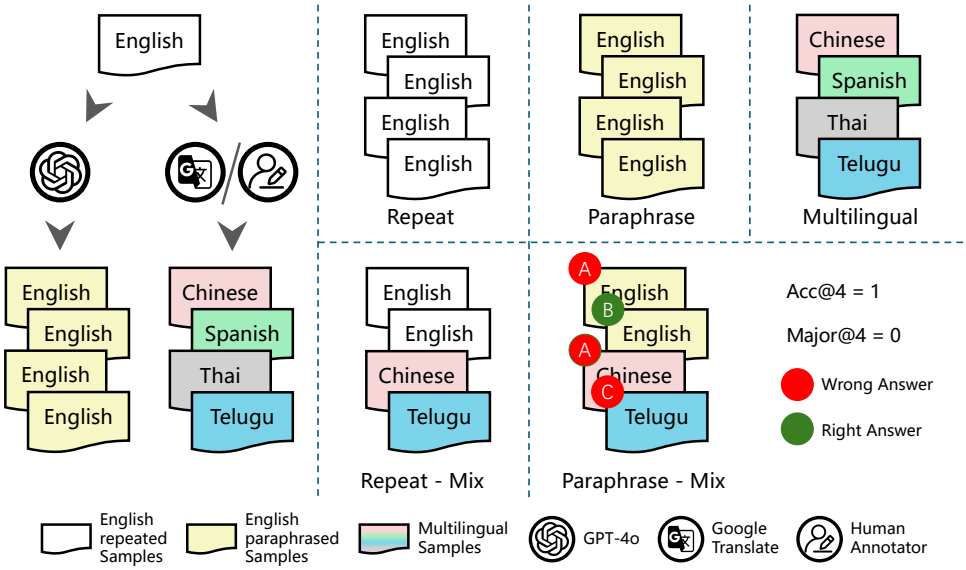

Figure 2: An introduction to input samples across various comparison methods, including Multilingual, Repeat, Paraphrase, Repeat-Mix, and Paraphrase-Mix.

and collect LLM responses:

- *Multilingual*: We translate the original English sample into various languages, and then prompt the model with each translation using a fixed random seed.
- *Repeat*: We repeatedly fed the model the same English sample, each time prompting it to generate a response using a different random seed.
- *Paraphrase*: We fed the model the paraphrased version of the original English sample generated by LLM, using a constant random seed.
- *Repeat-Mix*: We combine the Repeat and Multilingual responses in a $50/50$ split. One set of responses uses the original English text with different random seeds, while the other set uses translated texts with a consistent random seed. 2 out of 4 random seeds are used for the Repeat part and 2 out of 17 languages are used for the multilingual part, and the performances of all their combinations are collected.
- *Paraphrase-Mix*: We mixed the Paraphrase and Multilingual responses in equal parts $(50/50)$. One part is based on paraphrased versions, while the other part is derived from translated texts, with a fixed random seed used throughout. 2 out of 4 paraphrases are used for the Paraphrase part and 2 out of 17 languages are used for the multilingual part, and the performances of all their combinations are collected.

Models We use `Qwen2.5-72B`, `LLaMA3.1-70B` and `R1-Distill-LLaMA-70B` in our experiments. All results are based on their post-trained / instruction-tuned versions. We prompt these models to employ Chain-of-Thought (CoT) reasoning for all questions during inference. The prompt template is reported in Appendix A.

Testing Scenario We conduct our analysis on the Reasoning-Math and Reasoning-Science datasets drawn from BENCHMAX (Huang et al., 2025). The datasets are adapted from the GPQA (Rein et al., 2023) and MGSM (Shi et al., 2023) datasets, with human translations to support 17 languages. Reasoning-science is a multiple-choice task, while reasoning-math involves answering basic mathematical problems that require multi-step reasoning. In this paper, we primarily present the experimental results for reasoning-science, with the details of reasoning-math provided in Appendix B.

Metric The default metric that we used is Accuracy (Acc), which measures the agreement of the prediction generated by the model with the ground truth. $\overline{Acc}$ represents the average accuracy across k answer candidates. We use Acc@k metric to test the probability that at least one generated answer out of k for a problem is the ground truth. Major@k is utilized to assess model's accuracy after selecting answers from k candidates using a majority voting strategy. Judge@k is used in the LLM-as-a-judge experiments to denote the accuracy of the judged winners.

3.2 Intriguing Phenomena

We notice that multilingualism enhances reasoning, often resulting in surprising performance.

Phenomenon 1: Non-English languages can excel beyond English. Evaluating the performance of different models in reasoning-math and reasoning-science tasks reveals an intriguing phenomenon. Models trained on English-centric data do not always excel in reasoning tasks with the English language. This pattern is commonly observed across tasks, various model series, and models of different sizes, as shown in Figure 3.

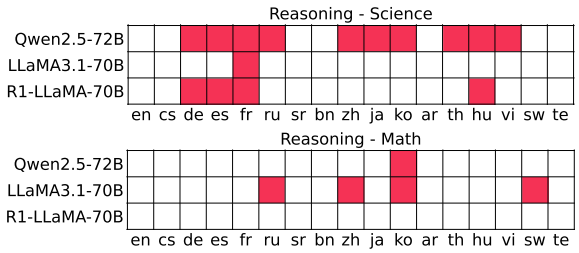

Figure 3: The cells highlighted in red indicate scores that are greater than the English scores in each row. English is not always better than other languages.

Phenomenon 2: Mixing languages boosts performance, setting higher upper-bound. Enhancing language diversity during model generation results in remarkable performance improvements, with the ceiling of these improvements notably high. Compared to the strategies of Repeat and Paraphrase, as depicted in Figure 4, the Multilingual can yield gains that far exceed 10 Acc@k points. Notably, while no individual non-English language in the combination outperforms English, their combination can still achieve significant improvements.

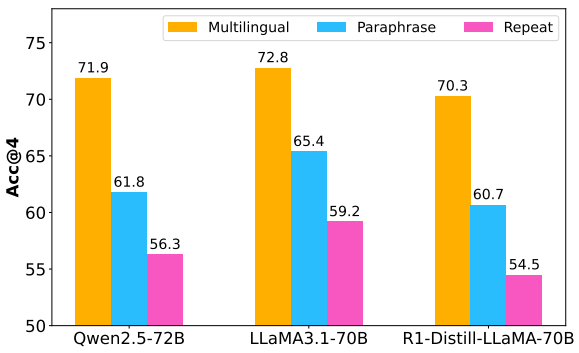

Figure 4: Compared to Repeat and Paraphrase, **Multilingual** demonstrates a higher upper-bound with Acc@k.

Phenomenon 3: A few languages combinations offer substantial performance boosts. In reasoning-science task, we rank 17 languages based on their performance in each model from high to low and combined the top-performing languages with varying numbers each time. As shown in Figure 5, as the number of mixed languages increases, the Acc@k performance consistently improves. Notably, just a few languages (2-4) can significantly enhance performance, quickly surpassing that of Repeat / Paraphrase.

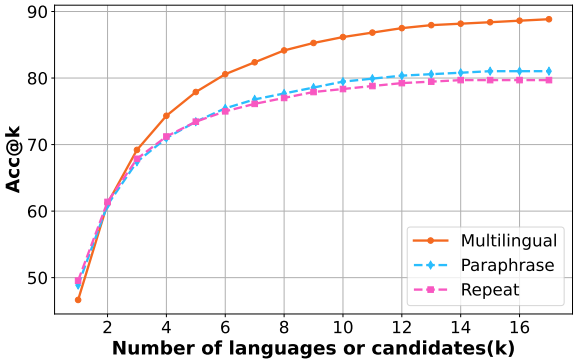

Figure 5: Best Acc@k for the tested models with increased numbers of languages or candidates on the Reasoning-Science task.

Phenomenon 4: Multilingual gain - Going beyond existing English benefits. As illustrated in Figure 6, multilingual input significantly enhances reasoning performance, surpassing the limits achieved by Repeat or Paraphrase. Notably, the improvements associated with multilingual input do not overlap with those derived from Repeat or Paraphrase methods. The experiments involving Repeat-Mix and Paraphrase-Mix indicate that replacing a portion of the input with multilingual data results in additional benefits to the performance upper-bound. This suggests that multilingual input provides a unique advantage in reasoning tasks, enabling models to leverage diverse linguistic structures and contexts.

Phenomenon 5: Upper-bound is tolerant of sub-optimal language choice. We evaluate all 4-combinations out of the 17 testing languages and observe that model performance significantly varies, as indicated by the low $\overline{Acc}$ in Table 1. Interestingly, the random scores show that, if the languages are randomly selected, the Acc@k score is still close to the best-performing language combinations.

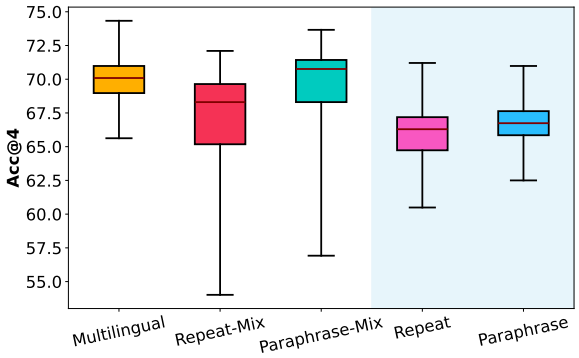

Figure 6: The Acc@4 score distribution of Qwen2.5-72B on the science reasoning task under different settings. Fully utilizing non-English languages can improve the upper-bound.

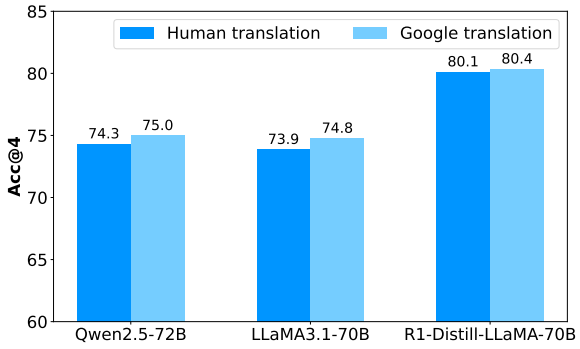

Figure 7: The Acc@k score of different models under Multilingual setting is stable regardless of the question translation quality, by comparing its performance with human annotations versus Google translate results.

Model	Setting	$\overline{Acc}$	Acc@4
Qwen2.5-72B	Best	43.7	74.3
	Worst	37.8	65.6
	Random	41.5	70.0
LLaMA3.1-70B	Best	38.0	73.9
	Worst	32.6	65.2
	Random	36.9	70.2
R1-Distill-LLaMA-70B	Best	51.6	80.1
	Worst	34.0	64.7
	Random	49.0	75.5

Table 1: Evaluation on reasoning tasks shows significant performance diversity, highlighted by low $\overline{Acc}$ across languages. However, the choice of languages demonstrates minimal influence on Acc@k, indicating a negligible impact on multilingual potential.

Phenomenon 6: Upper-bound is robust to translation quality. A high-quality multilingual dataset annotated by humans across multiple languages poses a challenge for many tasks. Therefore, we explore the quality of multilingual text further, investigating whether its superior translation quality is crucial for model performance. As depicted in Figure 7, there is a slight performance difference between the model outputs generated with human annotations and those from Google Translate. This experiment highlights that we can easily accessible multilingual capabilities can significantly enhance reasoning abilities.

Phenomenon 7: Upper-bound is sensitive to answer selection strategy. We conduct a series of experiments across different models to evaluate the effectiveness of various answer selection strategies, specifically Repeat, Paraphrase, and Multilingual settings, as shown in Table 2. In the Acc@4 metric, the Multilingual approach outperforms the other strategies. However, when analyzing the results with $\overline{Acc}$ and Major@4 metrics, Multilingual does not achieve similarly favorable outcomes. This inconsistency suggests that while Multilingual strategies can be advantageous in certain contexts, their overall effectiveness may be limited by the answer selection criteria employed. These findings underscore the necessity of different answer selection strategies across various models.

Model	Setting	Acc@4	$\overline{Acc}$	Major@4
Qwen2.5-72B	Repeat	71.2	**48.1**	53.7
	Paragraph	71.0	47.3	**54.4**
	Multilingual	**74.3**	43.7	54.2
LLaMA3.1-70B	Repeat	71.0	42.4	50.4
	Paragraph	73.0	**43.8**	**51.3**
	Multilingual	**73.9**	38.0	49.8
R1-Distill-LLaMA-70B	Repeat	77.9	**54.7**	**63.2**
	Paragraph	74.8	51.3	60.8
	Multilingual	**80.1**	51.6	61.2

Table 2: Performance comparison across different methods—Multilingual, Repeat, and Paraphrase—is conducted using Acc@4, $\overline{Acc}$, and Major@4 for answer selection. The discrepancies observed among these metrics highlight the critical role of answer selection in effectively leveraging multilingualism for reasoning.

4 Answer Selection Strategies

In this section, we explore other commonly used answer selection methods: Prompt-based Aggregation (§ 4.1), and LLM-as-a-judge selection (§ 4.2).

4.1 Prompt-based language selection

One straightforward method for answer selection is the prompt-based approach, which entails furnishing precise input instructions to direct the model in producing the desired outputs. To steer the model

towards maximizing its multilingual capabilities, we have customized prompts for the LLM from three crucial viewpoints: language constraint, English allowance, and question translation.

- Language Constraint (LC) provides a predefined set of languages that the model can utilize, guaranteeing a high-quality language combination and optimal performance.
- English Allowance (EA) relates to whether to incorporate English as one of the languages that can be used.
- Question Translation (QT) aims to explicitly prompt the model to leverage its multilingual capabilities through translation, encouraging the use of multiple languages in crafting responses.

Model	LC	EA	QT	Setting	Acc@4	Major@4
Qwen2.5-72B	-	-	-	Repeat	61.8	52.3
	-	-	-	Paraphrase	**68.3**	**53.9**
	✓	✗	✓	-	59.2	48.2
	✓	✓	✓	-	63.8	51.8
	✗	✓	✓	-	61.2	53.2
	✓	✗	✗	-	62.7	50.6
	✓	✓	✗	-	61.2	52.5
	✗	✓	✗	-	62.1	52.0
LLaMA3.1-70B	-	-	-	Repeat	61.2	49.3
	-	-	-	Paraphrase	**71.2**	**51.5**
	✓	✗	✓	-	58.9	46.6
	✓	✓	✓	-	61.8	47.5
	✗	✓	✓	-	65.6	50.1
	✓	✗	✗	-	62.5	46.6
	✓	✓	✗	-	63.2	50.6
	✗	✓	✗	-	65.0	49.6
R1-Distill-LLaMA-70B	-	-	-	Repeat	74.1	62.2
	-	-	-	Paraphrase	71.4	61.8
	✓	✗	✓	-	75.9	64.9
	✓	✓	✓	-	73.2	59.2
	✗	✓	✓	-	72.8	58.7
	✓	✗	✗	-	**76.8**	**66.3**
	✓	✓	✗	-	72.8	56.8
	✗	✓	✗	-	72.8	57.7

Table 3: In the reasoning-science task, the performance (Acc@4 and Major@4) on various prompt-based methods shows little variation between different approaches. The requirement of translation is not the key to multilingual reasoning, and no prompt-based method stands out. Here, the results of Repeat and Paraphrase involve averaging the scores across different 4 runs.

Prompt-based methods cannot unlock a model's multilingual capabilities during reasoning. As shown in Table 3, no prompt-based approach stands out as superior, with minimal differences between methods. However, there are still some intriguing discoveries:1) Translating the original question from English to non-English before responding does not affect the model's final performance. 2) Interestingly, when comparing LLaMA3.1-70B with R1-Distill-Llama-70B, prompt-based methods achieve better results than Repeat and Paraphrase.

Model	Setting	Acc@4	Major@4	Judge@4
Qwen2.5-72B	Repeat	61.4	53.4	48.9
	Paragraph	63.0	**54.2**	**50.4**
	Multilingual	**66.7**	51.4	43.1
LLaMA3.1-70B	Repeat	62.1	50.6	**47.1**
	Paragraph	65.8	49.2	46.2
	Multilingual	**67.6**	**50.9**	41.5
R1-Distill-LLaMA-70B	Repeat	71.2	57.2	57.1
	Paragraph	71.9	59.2	58.9
	Multilingual	**76.1**	**62.6**	**62.7**

Table 4: LLM-as-a-judge performance on the reasoning-science dataset.

4.2 LLM-as-a-judge selection

Another commonly used answer selection method is LLM-as-a-judge (Li et al., 2024), where a judge model evaluates two answers to a given question, and selects the best of them as the winner. Here, we use the tested models to judge their own outputs, and conduct pairwise judgments for each two of the candidates with position swapping, and take the one winning the most battles.

To test the effectiveness of this method, we run judges on the machine-translated multilingual questions, using the best language combination found on them for each model and collect the accuracies of the judged outputs (Judge@k). Then, we compare them with English Repeat and Paraphrase with the same judging process.

The results are shown in Table 4. Still, while Multilingual leads the Acc@k scores, its Judge@k scores are lower than the English baselines except for the R1-Distill-LLaMA model. Also, the Judge@k scores in most of the tested settings are lower than Major@k scores, meaning LLM-as-a-judge is even less effective than simple majority voting in answer selection. This suggests LLM-as-a-judge answer selection can be biased.

5 Analysis

While §3 shows the high upper-bound gain of multilingualism, §4 shows that common selection approaches have difficulty realizing this gain. Here, we discuss the reasons behind this gap.

5.1 Possible reasons for the upper-bound gain

We propose several possible reasons for the upper-bound gain of multilingualism.

Language correctness correlates with question difficulty. The first hypothesis is that different languages match questions of different levels of difficulty. Simple questions, such as those about commonly-known science knowledge or simple

math calculation, are frequently seen in the general-domain data, making English a suitable language for the models to use. However, for hard questions that require much domain expertise, it will be beneficial to use specific languages that cover similar domain-specific knowledge in the training data.

To verify this hypothesis, on the reasoning-science tasks where question difficulties are labeled, we calculate the difficulty distribution of the correctly answered questions in each language. The results (Table 5) show that, the difficulties of the correctly answered questions varies in a small range across languages. For the easiest and the hardest questions, this varying difficulty distribution is more evident.

Model		Easy Undergrad	Hard Undergrad	Hard Grad	Post-Grad
Qwen2.5-72B	Max	1.7	58.7	36.7	10.8
	Min	0.5	53.6	30.7	6.5
LLaMA3.1-70B	Max	1.5	56.5	38.6	11.4
	Min	0.5	49.1	31.6	6.5
R1-Distill-LLaMA-70B	Max	1.4	59.8	33.6	9.2
	Min	0.5	54.8	26.3	6.6

Table 5: Difficulty distributions across languages of the correctly answered questions.

Existence of key advantaged languages The last hypothesis is that, for a model on a specific task with multilingual reasoning, there will be some key advantage languages that often compensate errors in other languages, which contributes to the high Acc@k. Furthermore, if the key advantaged languages overlap on different models, it will be likely that these languages are more suitable than others on the specific task.

We set a standard called the minority-majority overlap to identify such language advantage. First, we collect the languages with high accuracies, both on questions correctly answered only in a few languages, and by a vast majority of the languages Then, we report the overlap of the leading languages in the both situations. Finally, we report the cross-model overlap of these languages. As shown in Table 6, each model has some key advantaged languages in the two tasks, respectively, and there are also cross-model key advantage languages, namely French for reasoning-science, and Korean and English for reasoning-math.

5.2 Challenges to meet the upper-bound

We will discuss some challenges in meeting the multilingual reasoning upper-bound with common approaches.

Task	Model	Advantage Langs
Science	Qwen2.5-72B	ja,en,fr,hu
	LLaMA3.1-70B	hu,en,fr,ru,de
	R1-Distill-LLaMA-70B	es,vi,cs,fr
Math	Qwen2.5-72B	ko,ar,es,en,sr,vi,hu
	LLaMA3.1-70B	ru,ko,en,es,vi,de
	R1-Distill-LLaMA-70B	sr,ar,ko,en,cs,hu

Table 6: Key advantaged language found by minority-majority overlap.

Voting performance does not grow with language numbers. As shown in Figure 8, as the size of the language combination grows up, the Major@k score does not increase, but declines instead, which is the opposite from the Acc@k curve in Figure 5. This is mainly because the gain and advantage of multilingualism in Acc@k is often brought by only a few languages, especially when the majority is wrong. Thus, a larger number of languages can bring more noise, making it harder for the correct answer to win majority.

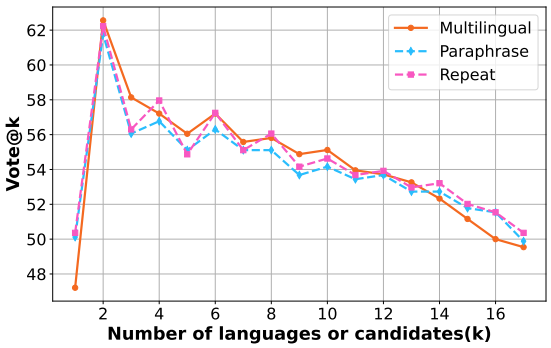

Figure 8: Best Major@k for the tested models with increased numbers of languages or candidates on the Reasoning-Science task

Voting performance relies on optimal language combination. While we show the multilingual reasoning upper-bound is tolerant to sub-optimal language combinations in §3.2, the multilingual majority voting performance relies on optimal language combinations to surpass English voting. As shown in Figure 9, the voting accuracy of Multilingual is higher than or quite close to the those of Paraphrase and Repeat if all of them use their best language combinations. However, when the language combination is random or the worst, the Multilingual voting accuracy will be lower than the other two, indicating that majority voting on the

Model	LC	EA	QT	En	Max Non-En
Qwen2.5-72B	✓	✗	✓	4.4	45.5
	✓	✓	✓	99.9	0.1
	✗	✓	✓	99.7	0.3
	✓	✗	✗	62.1	17.2
	✓	✓	✗	99.8	0.2
	✗	✓	✗	99.8	0.2
LLaMA3.1-70B	✓	✗	✓	1.1	83.4
	✓	✓	✓	46.5	53.3
	✗	✓	✓	99.7	0.2
	✓	✗	✗	25.6	52.5
	✓	✓	✗	85.6	14.1
	✗	✓	✗	99.9	0.1
R1-Distill-LLaMA-70B	✓	✗	✓	100.0	0.0
	✓	✓	✓	99.9	0.1
	✗	✓	✓	99.9	0.1
	✓	✗	✗	99.9	0.1
	✓	✓	✗	99.9	0.1
	✗	✓	✗	99.8	0.1

Table 7: Language chosen rate of the prompt-based answer selection methods. We report the chosen rate of English and the highest non-English language.

Multilingual setting is sensitive to the optimality of the language combination.

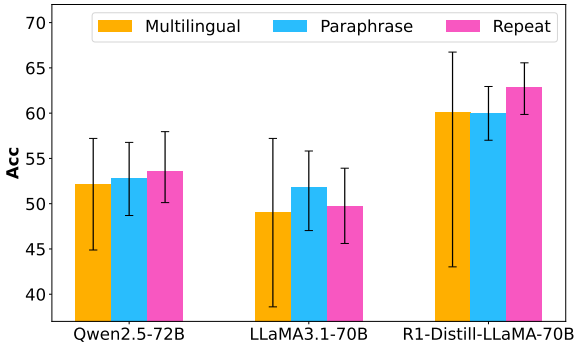

Figure 9: Comparison of Major@4 scores of Repeat, Paraphrase and Multilingual on the machine-translated reasoning-science dataset with random combinations of languages or runs. The error bars denotes the scores of the best and worst combinations.

Prompt-based and LLM-as-a-judge selection have language bias. For LLM-based answer selection methods, namely prompt-based and LLM-as-a-judge, an observed challenge is that they are biased to English and other high-resource language (e.g. Spanish, Japanese, etc.).

For prompt-based selection, the models tend to choose the high-resource languages for all the questions, thus decreasing the diversity of the candidate answers. Table 7 shows English and the most frequently chosen non-English languages and their rates in different settings, and the result show that, when English is allowed, the models will choose English in most cases; and when it is not allowed, the models tend to choose a certain language (such as Spanish or Vietnamese) than other languages in most cases.

Model	Lang	Chosen when correct	Chosen when incorrect
Qwen2.5-72B	ar	24.2	21.6
	de	31.4	31.9
	ja	11.3	13.8
	zh	35.5	31.0
LLaMA3.1-70B	de	35.6	25.9
	en	48.9	41.6
	hu	8.5	12.9
	ru	10.5	17.0
R1-Distill-LLaMA-70B	ar	20.8	12.9
	es	40.5	24.5
	ru	22.3	13.7
	vi	36.1	27.6

Table 8: The language chosen rates in LLM-as-a-judge based on whether the answers are correct or incorrect in these languages.

Similarly, for LLM-as-a-judge selection, the judge model tend to prefer answers in high-resource languages, even if the answer of that language is incorrect. Table 8 shows chosen rate of languages when the answer in that language is correct or incorrect. The results show that, except for the R1-Distill-LLaMA-70B model, the other two models show minor difference in chosen rate when the answer is correct or incorrect, suggesting that these two models care more for the language instead of the correctness of the answer while judging. This can also explain why the LLM-as-a-judge method only works for R1-Distill-LLaMA-70B.

6 Conclusion

In this paper, we comprehensively explore the benefits of multilingualism in reasoning and highlight several intriguing phenomena. Our findings suggest that utilizing multiple languages can significantly enhance reasoning capabilities, with a high upper bound for this benefit. Notably, this advantage is resilient to variations in translation quality and language choice, yet it remains sensitive to the methods used for answer selection. We examine various commonly used answer selection techniques but find that they often fall short of fully harnessing the potential of multilingualism in reasoning tasks. This disparity between the theoretical upper bound and practical experimental outcomes presents both a challenge and a promising avenue for future research.

Limitations

In this study, while providing valuable insights into the potential of multilingualism in reasoning, has several notable limitations. However, our focus is primarily on large models with over 70 billion parameters, which may not fully represent the capabilities or challenges faced by smaller models. This narrow scope could lead to an incomplete understanding of how multilingualism affects reasoning across various architectures and sizes. Additionally, although we observe several interesting phenomena, the absence of a universal and stable method for leveraging multilingualism in reasoning.

References

Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. To code, or not to code? exploring impact of code in pre-training. *arXiv preprint arXiv:2408.10914*.

Dörthe Bühmann. 2008. *Mother tongue matters: Local language as a key to effective learning*. Unesco.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, and Pengfei Liu. 2022. Polyglot prompt: Multilingual multitask prompt training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9919–9935, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gemini. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Lianzhe Huang, Shuming Ma, Dongdong Zhang, Furu Wei, and Houfeng Wang. 2022. Zero-shot cross-lingual transfer of prompt-based tuning with a unified multilingual prompt. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11488–11497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *Preprint*, arXiv:2502.07346.

Junlong Li, Daya Guo, Dejian Yang, Runxin Xu, Yu Wu, and Junxian He. 2025. Codei/o: Condensing reasoning patterns via code input-output prediction. *Preprint*, arXiv:2502.07316.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *Preprint*, arXiv:2406.11939.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue,

Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Qwen Team. 2025. Qwen2.5 technical report.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

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

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

DeepSeek-AI Team. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik R Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Fei Yuan, Shuai Yuan, Zhiyong Wu, and Lei Li. 2024. How vocabulary sharing facilitates multilingualism in LLaMA? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12111–12130, Bangkok, Thailand. Association for Computational Linguistics.

Wenhao Zhu, Shujian Huang, Fei Yuan, Cheng Chen, Jiajun Chen, and Alexandra Birch. 2024. The power of question translation training in multilingual reasoning: Broadened scope and deepened insights. *arXiv preprint arXiv:2405.01345*.

A Models

A.1 Model Description

Qwen2.5-72B is a cutting-edge language model designed to enhance natural language processing tasks with its impressive 72 billion parameters. This model excels in generating coherent and contextually relevant text, making it particularly valuable for applications in content creation, conversational agents, and automated summarization.

LLaMA3.1-70B represents the latest iteration in the LLaMA series, boasting 70 billion parameters that empower it to tackle complex reasoning tasks and generate high-quality text. This model is particularly noted for its ability to engage in multi-turn conversations, maintaining context and coherence over extended interactions.

R1-Distill-LLaMA-70B is a distilled version of the original LLaMA model, optimized for efficiency without compromising performance. With 70 billion parameters, this model is designed to deliver faster response times and reduced computational requirements, making it ideal for deployment in resource-constrained environments.

A.2 Languages-Related Prompt

We present the prompt templates utilized in our experiments, including the Default and Prompt-based selection, as shown in Table 9. In the prompt-based selection experiments, we incorporated language-related constraints regarding whether to translate the question. Consequently, there are two variations of prompt-based selection: Translation=True and Translation=False, as indicated in the table.

B Results on Reasoning-Math

We demonstrate the results of the three models on the reasoning-math task. The results of the Repeat, Paraphrase, Multilingual, Repeat-Mix, and Paraphrase-Mix methods are presented in Table 10, Table 11, Table 12, Table 13, and Table 14, respectively. Table 15 shows the results on the Google translated reasoning-math task.

C 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.

- *LLaMA-3.1 (LLaMA3.1 license)*, a large language model developed by Meta.
- *R1-Distill-LLaMA-70B (MIT license)*, a large language model developed by Deepseek.
- *Qwen-2.5-72B (Qwen license)*, a large language model developed by Qwen.

Setting	Prompt
	Reasoning-Science
Default	**System prompt:** Always think step by step and give your final choice among (A), (B), (C) and (D) by Änswer: {Your Choice}ïn a single last line. **User prompt:** What is the correct answer to this question:{Question} Choices: (A) choice1 (B) choice2 (C) choice3 (D) choice4 Let's think step by step:
Prompt-Based Selection Translation = True	**System prompt:** Always choose the most suitable language, translate the question into that language, and think step by step in that language. Give your final choice among (A), (B), (C) and (D) by Änswer: {Your Choice}ïn a single last line. **User prompt:** What is the correct answer to this question:{Question} Choices: (A) choice1 (B) choice2 (C) choice3 (D) choice4 Let's think step by step:
Prompt-Based Selection Translation = False	**System prompt:** Always choose the most suitable language, and think step by step in that language. Give your final choice among (A), (B), (C) and (D) by Änswer: {Your Choice}ïn a single last line. **User prompt:** What is the correct answer to this question:{Question} Choices: (A) choice1 (B) choice2 (C) choice3 (D) choice4 Let's think step by step:
	Reasoning-Math
Default	**System prompt:** Always think step by step and give your final answer by Änswer: Your Answerïn a single last line. **User prompt:** Question: {Question} Step-by-Step Answer:
Prompt-Based Selection Translation = True	**System prompt:** Always choose the most suitable language, translate the question into that language, and think step by step in that language. Give your final answer by Änswer: Your Answerïn a single last line. **User prompt:** Question: {Question} Step-by-Step Answer:
Prompt-Based Selection Translation = False	**System prompt:** Always choose the most suitable language, and think step by step in that language. Give your final answer by Änswer: Your Answerïn a single last line. **User prompt:** Question: {Question} Step-by-Step Answer:

Table 9: The prompt template we used in experiments for each task.

Model	Setting	English Seed1	English Seed2	English Seed3	English Seed4	$\overline{\text{Acc}}$	Acc@4	Major@4
Qwen2.5-72B	Best	93.6	91.6	92.0	92.8	92.5	94.0	93.6
	Worst	91.6	92.0	91.2	92.0	91.7	92.4	92.0
	Random	-	-	-	-	92.4	93.7	92.5
LLaMA3.1-70B	Best	91.6	92.0	92.4	93.6	92.4	96.0	93.2
	Worst	90.0	90.8	91.6	91.6	91.0	92.4	91.2
	Random	-	-	-	-	91.7	94.8	92.0
R1-Distill-LLaMA-70B	Best	93.6	92.8	94.0	94.4	93.7	97.2	94.4
	Worst	93.6	94.0	91.6	92.0	92.8	94.0	94.0
	Random	-	-	-	-	93.6	96.0	94.0

Table 10: The results of the Repeat method on the reasoning-math task.

Model	Setting	English Paraphrase1	English Paraphrase2	English Paraphrase3	English Paraphrase4	$\overline{\text{Acc}}$	Acc@4	Major@4
Qwen2.5-72B	Best	92.0	88.8	90.8	91.6	90.8	96.0	93.2
	Worst	89.6	88.4	88.0	88.4	88.6	92.4	90.0
	Random	-	-	-	-	89.9	94.6	91.0
LLaMA3.1-70B	Best	90.0	90.0	91.2	88.8	90.0	96.4	91.6
	Worst	86.0	88.4	87.2	88.8	87.6	92.4	89.2
	Random	-	-	-	-	88.9	94.8	90.8
R1-Distill-LLaMA-70B	Best	89.6	91.2	91.2	90.4	90.6	96.8	91.6
	Worst	89.6	89.2	89.2	89.2	89.3	92.8	92.0
	Random	-	-	-	-	90.0	94.9	91.4

Table 11: The results of the Paraphrase method on the reasoning-math task.

Model	Setting	Lang-1	Lang-2	Lang-3	Lang-4	$\overline{\text{Acc}}$	Acc@4	Major@4
Qwen2.5-72B	Best (ar,cs,en,ko)	91.2	84.8	92.4	91.2	89.9	98.0	94.0
	Worst (bn,sw,te,zh)	83.2	63.2	62.8	86.4	73.9	92.8	86.8
	Random	-	-	-	-	84.7	95.8	91.7
LLaMA3.1-70B	Best (ar,ru,sr,vi)	87.2	90.0	87.2	90.0	88.6	99.6	92.8
	Worst (bn,cs,sw,zh)	81.2	85.2	84.4	84.4	83.8	93.6	90.
	Random	-	-	-	-	86.5	96.9	92.0
R1-Distill-LLaMA-70B	Best (ar,bn,de,sr)	90.8	75.2	86.4	92.4	86.2	99.6	92.8
	Worst (bn,de,te,vi)	75.2	86.4	78.0	87.6	81.8	95.6	90.4
	Random	-	-	-	-	86.4	97.8	92.6

Table 12: The results of the Multilingual method on the reasoning-math task.

Model	Setting	Lang-1	Lang-2	Lang-3	Lang-4	$\overline{\text{Acc}}$	Acc@4	Major@4
Qwen2.5-72B	Best (en,en,es,te)	92.4	92.8	90.8	62.8	84.7	97.6	93.2
	Worst (en,en,ar,bn)	92.4	92.8	91.2	83.2	89.9	93.2	92.8
	Random	-	-	-	-	85.8	95.4	92.8
LLaMA3.1-70B	Best (en,en,es,th)	91.6	91.6	92.0	87.2	90.3	99.2	92.0
	Worst (en,en,cs,de)	91.6	91.6	85.2	88.8	89.3	92.8	91.2
	Random	-	-	-	-	87.2	96.9	92.5
R1-Distill-LLaMA-70B	Best (en,en,es,vi)	92.8	93.2	87.2	87.6	90.2	99.6	94.0
	Worst (en,en,ar,bn)	92.8	93.2	90.8	75.2	88.0	94.8	93.2
	Random	-	-	-	-	87.5	97.6	94.1

Table 13: The results of the Repeat-Mix method on the reasoning-math task.

Model	Setting	Lang-1	Lang-2	Lang-3	Lang-4	$\overline{\text{Acc}}$	Acc@4	Major@4
Qwen2.5-72B	Best (en,en,es,te)	92.0	88.8	90.8	62.8	83.6	98.4	91.6
	Worst (en,en,bn,cs)	88.8	89.6	83.2	84.8	86.6	92.0	91.2
	Random	-	-	-	-	85.4	95.7	91.8
LLaMA3.1-70B	Best (en,en,es,te)	90.0	90.0	92.0	82.8	88.7	99.2	93.6
	Worst (en,en,ar,cs)	90.0	90.0	87.2	85.2	88.1	92.8	91.6
	Random	-	-	-	-	86.9	96.8	92.7
R1-Distill-LLaMA-70B	Best (en,en,es,vi)	89.6	91.6	87.2	87.6	89.0	99.6	93.2
	Worst (en,en,cs,de)	89.6	89.2	87.2	86.4	88.1	92.0	92.4
	Random	-	-	-	-	86.8	96.7	92.7

Table 14: The results of the Paraphrase-Mix method on the reasoning-math task.

Model	Setting	Lang-1	Lang-2	Lang-3	Lang-4	$\overline{\text{Acc}}$	Acc@4	Major@4
Qwen2.5-72B	Best (ar,en,es,hu)	88.8	92.4	89.2	78.8	87.3	98.0	92.4
	Worst (bn,cs,fr,sw)	29.2	82.4	82.4	62.8	64.2	92.4	84.0
	Human (ar,cs,en,ko)	88.8	82.4	92.4	88.0	87.9	96.4	92.4
	Random	-	-	-	-	80.5	96.0	90.7
LLaMA3.1-70B	Best (de,fr,ja,vi)	88.0	82.8	83.6	89.2	85.9	99.2	90.4
	Worst (bn,cs,sw,zh)	81.6	84.0	79.6	84.8	82.5	94.0	90.0
	Human (ar,ru,sr,vi)	86.4	89.2	85.6	89.2	87.6	97.2	93.6
	Random	-	-	-	-	85.0	96.8	91.4
R1-Distill-LLaMA-70B	Best (ar,hu,sr,zh)	89.6	82.0	87.2	88.4	86.8	98.8	94.0
	Worst (de,sw,te,vi)	86.8	80.0	77.2	84.8	82.2	93.6	91.2
	Human (ar,bn,de,sr)	89.6	52.8	86.8	87.2	79.1	98.0	92.8
	Random	-	-	-	-	83.6	97.2	92.1

Table 15: The results of the Multilingual method on the Google translated reasoning-math task.

Model	Langs	Acc@k	Major@k	Judge@k
Qwen2.5-72B	repeat	93.6	93.2	91.2
	paraphrase	94.0	91.6	90.0
	ar,en,es,hu	98.0	92.4	89.6
LLaMA3.1-70B	repeat	94.8	92.0	92.0
	paraphrase	95.2	91.6	91.6
	de,fr,ja,vi	99.2	90.4	88.0
R1-Distill-LLaMA-70B	repeat	96.4	93.6	92.8
	paraphrase	93.6	91.6	88.8
	ar,hu,sr,zh	98.8	94.0	91.6

Table 16: LLM-as-a-judge performance on reasoning-math dataset