How Well Does Mathematical Reasoning Transfer Across Languages in English-Centric LLMs?

Anonymous ACL submission

Abstract

Recent advances in Reinforcement Post-Training (RPT) have substantially improved the mathematical reasoning capabilities of large language models (LLMs), but it remains unclear how well such gains transfer across languages. In this work, we study cross-lingual mathematical reasoning transfer in English-centric LLMs under controlled multilingual evaluation. We first evaluate a range of English-centric reasoning models on multilingual mathematical reasoning benchmarks and find that transfer is highly variable across initial model choice, target language, and training paradigm. Through controlled comparative experiments, we further find that stronger English-centric initialization does not necessarily lead to stronger relative transfer efficiency, even when it yields higher multilingual reasoning accuracy. Finally, parallel multilingual supervision improves transfer in our setting, with the largest observed single-step gain from English-only to bilingual training and smaller positive gains from additional languages. Overall, our results highlight the need to evaluate multilingual mathematical reasoning directly, using both absolute performance and relative transfer-efficiency metrics.

1 Introduction

Recent advances in Reinforcement Post-Training (RPT) (Jaech et al., 2024; Kimi et al., 2025; Team, 2025) have substantially improved the reasoning capabilities of large language models (LLMs), especially on mathematical benchmarks. In particular, Reinforcement Learning with Verifiable Rewards (RLVR) (Lambert et al., 2024; Guo et al., 2025) has led to strong gains on datasets such as MATH (Hendrycks et al., 2021) and AIME. As these models continue to improve on English reasoning benchmarks, an important question is whether such gains transfer beyond English. Prior work has mainly studied transfer across tasks or modalities (Chu et al., 2025; Liu et al., 2025a; Hu et al., 2025a; Huan et al., 2025; Zhou et al., 2025), while cross-lingual transfer remains less understood.

This question matters for both multilingual use and model analysis. Strong English reasoning performance does not guarantee equally strong performance for multilingual users, especially when users require both correct answers and interpretable reasoning traces in their own language. More broadly, cross-lingual transfer offers a way to probe what English-centric post-training learns: transferable reasoning behavior, language-specific patterns, or some combination of both. This motivates the central question of this paper:

> ***(Q)*** *To what extent does mathematical reasoning acquired through English-centric post-training transfer across languages under controlled multilingual evaluation?*

We answer this question in three stages. We first conduct an observational study over open-source English-centric reasoning models and find that cross-lingual transfer is highly variable across initial model choice, target language, and training paradigm. We then run controlled comparative studies on initialization type, model family, and model size. Across these settings, stronger English-centric initialization does not necessarily correspond to larger relative cross-lingual transfer gains. Finally, we study parallel multilingual supervision and find that it substantially improves transfer, with the largest gain appearing when moving from monolingual to bilingual training.

Our contributions are threefold. First, we provide an exploratory empirical study of cross-lingual mathematical reasoning transfer in English-centric LLMs, using both absolute performance and relative transfer-efficiency metrics. Second, under controlled GRPO-based post-training, we show that transfer varies across initialization type, model fam-

ily, and model size. Third, we find that parallel multilingual supervision improves transfer in our setting, with the largest observed single-step gain from English-only to bilingual training and smaller but consistently positive gains from additional languages.

We organize the paper around the following research questions:

- ***RQ1:*** To what extent do English-centric LLMs transfer mathematical reasoning across languages? (Section 2)
- ***RQ2:*** What factors are associated with stronger or weaker cross-lingual transfer under controlled post-training settings? (Section 3)
- ***RQ3:*** How can cross-lingual transfer be improved effectively? (Section 4)

2 Observational Study

To answer ***RQ1***—*To what extent do English-centric LLMs transfer mathematical reasoning across languages?*—we begin with an observational study of open-source reasoning models. This analysis establishes whether cross-lingual transfer is present and examines how it varies with model initialization, target language, and training paradigm.

2.1 Setup

Models. We evaluate open-source reasoning models tuned with either SFT or RLVR, including the Simple-Zoo (Zeng et al., 2025), s1 (Muennighoff et al., 2025), OpenThinker (Guha et al., 2025), Open-Reasoner-Zero (Hu et al., 2025b), and DeepSeek-R1-Distill (Guo et al., 2025) series.

Benchmarks. We evaluate on multilingual mathematical reasoning benchmarks, including the multilingual versions of MATH500 (Hendrycks et al., 2021), AIME2024, and AIME2025 from XReasoning (Qi et al., 2025). We also include multilingual GPQA-Diamond from BenchMAX (Huang et al., 2025b) as an out-of-domain reasoning benchmark. The evaluation includes *English(En), Spanish (es), Russian (ru), German (de), French (fr), Bengali (bn), Swahili (sw), Thai (th), Japanese (ja), Chinese (zh), and Telugu (te)*—yielding eleven evaluation languages in total. More details are given in Appendix C.1.

Language-controlled prompting. Following prior work (Luo et al., 2025; Wang et al., 2025b; Qi et al., 2025), we concatenate the language control instructions after the input prompts to make the model generate responses using the same language as the query. This setup reflects multilingual application scenarios where both the answer and the reasoning trace should be understandable in the user's language (Yong et al., 2025; Wang et al., 2025a; Zhang et al., 2025). The full prompts are provided in Appendix H.2 and H.3.

Metrics. We report reasoning accuracy (**Acc**) as the main task metric, and off-target rate (**Off-target**) to measure how often a model fails to respond in the requested language, using the LangDetect library. In all evaluation settings, correctness is determined from the final answer only. Intermediate reasoning traces are not scored. Therefore, a model does not receive higher accuracy simply for producing its reasoning trace in the target language.

To quantify relative transfer, we also use the *Multilingual Transferability Index* (MTI), following prior work (Huan et al., 2025; Huang et al., 2025a). MTI is intended to measure how much of the gain obtained in the training language transfers to unseen languages. Since base difficulty varies across languages and benchmarks, absolute accuracy alone is not sufficient for this purpose.

Let $Acc_{b,l}^{\text{trained}}$ and $Acc_{b,l}^{\text{base}}$ denote the accuracy of a trained model and its corresponding base model on benchmark b and language l. We define the relative gain on language l as

$$\Delta R_{b,l} = \frac{Acc_{b,l}^{\text{trained}} - Acc_{b,l}^{\text{base}}}{Acc_{b,l}^{\text{base}}}. \quad (1)$$

For a training language set $\mathcal{L}_{\text{train}}$, the average relative gain is

$$\Delta R_{b,\mathcal{L}_{\text{train}}} = \frac{1}{|\mathcal{L}_{\text{train}}|} \sum_{l \in \mathcal{L}_{\text{train}}} \Delta R_{b,l}. \quad (2)$$

We define the MTI for an unseen language l_{unseen} as the average benchmark-level relative transfer ratio:

$$\text{MTI}_{l_{\text{unseen}}} = \frac{1}{|B|} \sum_{b \in B} \frac{\Delta R_{b,l_{\text{unseen}}}}{\Delta R_{b,\mathcal{L}_{\text{train}}}}. \quad (3)$$

An MTI of 1 indicates that the relative gain on an unseen language matches the average gain on the training language(s); values below 1 indicate weaker relative transfer, and values above 1 indicate stronger relative transfer. MTI measures relative transfer efficiency rather than absolute multilingual reasoning ability: it captures how much

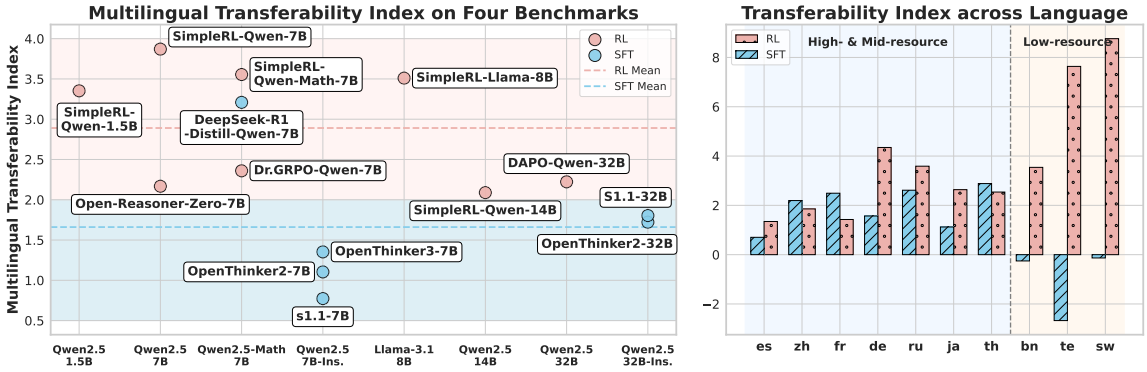


Figure 1: **Cross-lingual transfer across open-source reasoning models.** The *left* panel reports average MTI across four benchmarks and eleven languages. The *right* panel shows average transfer behavior on multilingual MATH500, comparing SFT-tuned and RL-tuned models.

of the training-language gain is reflected in unseen languages. Thus, a high MTI can still correspond to low absolute accuracy, especially from a low base accuracy. For this reason, we use MTI only as an auxiliary diagnostic and interpret it together with absolute accuracy, non-English average accuracy, low-resource accuracy, and off-target rate throughout the paper, detailed in Appendix E.2.

As a robustness check in Appendix F.1, we also repeated the main analyses with alternative transfer metrics, including normalized gain and raw gain ratio, and examined the sensitivity of MTI to low-base-accuracy cases. The numerical values differ, but the main qualitative conclusions remain unchanged.

2.2 Main Results

The main result of this section is that *cross-lingual transfer of mathematical reasoning is highly variable*. Figure 1 shows that strong English reasoning does not reliably translate into equally strong multilingual performance, and transfer varies across model initialization, training paradigm, and target language.

Model choice matters. Transfer varies substantially across models. Even among systems with similar training goals, multilingual performance differs markedly across languages. For example, SimpleRL-Qwen-7B shows slightly higher transferability than SimpleRL-Qwen-Math-7B despite similar training procedures.

English performance is not enough. High English performance and strong multilingual transfer are related, but not identical. Some models that perform well in English do not maintain equally strong multilingual performance, while others with weaker English starting points show larger relative gains after post-training. English evaluation alone is therefore insufficient for characterizing multilingual transfer.

Training paradigm matters. In the open-source model comparison, RL-tuned models generally show stronger multilingual performance than SFT-tuned models, especially in lower-resource languages such as *bn*, *sw*, and *te*. In these languages, SFT-tuned models often show weak or negative transfer, whereas RL-tuned models more often yield positive gains. Because the compared models also differ in initialization, training data, and implementation details, we treat this as an observational pattern rather than a controlled comparison of SFT and RL. A matched SFT-vs.-GRPO comparison on Qwen2.5-7B-Instruct is provided in Appendix E.1.

Target language matters. Transfer is not uniform across languages. Higher-performing languages are generally more stable, whereas lower-performing languages, including several low-resource languages, show larger relative variation across models and training paradigms. This variation may partly reflect higher instability in low-accuracy regimes, rather than linguistic factors alone. Overall, cross-lingual transfer depends not only on English-side training, but also on target-language difficulty and response-language stability.

3 Controlled Comparative Study

The observational study shows that cross-lingual transfer is highly variable, but open-source model comparisons cannot isolate the sources of this variation because models differ in initialization, data, optimization, and model family. To answer **RQ2**, we conduct controlled studies of initializa-

tion type, model family, and model size under matched GRPO-based post-training.

3.1 Setup

Training data. We use a compact training set of 1,000 examples sampled from the MATH training split, following prior work on small but carefully designed reasoning supervision sets (Ye et al., 2025; Muennighoff et al., 2025). All controlled studies in this section use the same training set unless otherwise noted. Details are provided in Appendix D.2.

Post-training algorithm. We use Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as our RPT algorithm. Following prior work (Rastogi et al., 2025; Zhang et al., 2025; Liu et al., 2026), the policy is optimized with a composite reward that combines reasoning accuracy, response format, and language consistency:

$$R = \lambda_1 R_{\text{acc}} + \lambda_2 R_{\text{format}} + \lambda_3 R_{\text{lang}}. \quad (4)$$

Here, R_{acc} rewards correctness, R_{format} encourages well-formed reasoning traces, and R_{lang} encourages responses in the target language. Hyperparameter details are given in Appendix D.4 and F.3.1.

Because the objective includes formatting and language-consistency terms, gains should not be interpreted as pure reasoning gains alone; we keep the objective fixed across settings for controlled comparison.

Controlled factors. We study three factors: *initial model type*, *model family*, and *model size*. In each case, all training data and optimization settings are fixed.

3.2 Initial Model Type

We first study the effect of initial model type using three Qwen2.5-7B variants: a general base model, an instruction-tuned model, and a math-specialized model. Table 1 shows a trade-off between final multilingual performance and relative transfer efficiency. The instruction model achieves the highest multilingual accuracy after English-only post-training and the lowest off-target rate, but the base and math-specialized models obtain higher MTI. This suggests that stronger final multilingual performance and stronger relative transfer efficiency are not always aligned under our setup. Notably, this observation concerns relative transfer efficiency rather than final multilingual performance, which remains highest for the instruction-tuned model under our setup.

Model	Acc	Off-target	MTI
Qwen2.5-7B-Base	12.00	11.41	-
↳ *GRPO on En Data*	22.45	3.12	1.95
Qwen2.5-7B-Instruct	22.45	1.43	-
↳ *GRPO on En Data*	23.51	0.94	1.23
Qwen2.5-Math-7B	12.25	22.59	-
↳ *GRPO on En Data*	19.37	9.50	2.12

Table 1: **The Impact of Initial Model Type.** Average Accuracy (%), Off-target rate (%), and MTI across eleven languages on four benchmarks.

3.3 Model Family

We next compare Qwen2.5-7B-Instruct and Llama3.1-8B-Instruct under the same English-only GRPO setup to test whether the pattern above persists across model families.

Figure 2 shows that English-only GRPO improves performance in both English and unseen languages for both families. However, the relative gains differ substantially: although Llama3.1-8B-Instruct starts from weaker English performance than Qwen2.5-7B-Instruct, it exhibits larger relative gains across languages. This suggests that cross-lingual transfer is not simply determined by English performance alone. At the same time, this comparison does not isolate a pure family effect, since differences in tokenization, pretraining data, and architecture remain entangled. Notably, this comparison concerns relative gains rather than final multilingual accuracy, and should be interpreted accordingly.

3.4 Model Size

We finally compare Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct under the same English-only GRPO setup to study how the observed pattern changes with model scale.

Figure 3 shows that the effect of model scale is benchmark-dependent. On multilingual MATH500, the 1.5B model gains more than the 7B model, suggesting greater headroom on the in-domain task. On multilingual AIME24/25, however, the 7B model attains stronger final performance on the harder benchmarks despite smaller gains on MATH500. On multilingual GPQA-Diamond, the smaller model again shows larger improvements, while the 7B model improves only marginally and can even degrade in English. One possible expla-

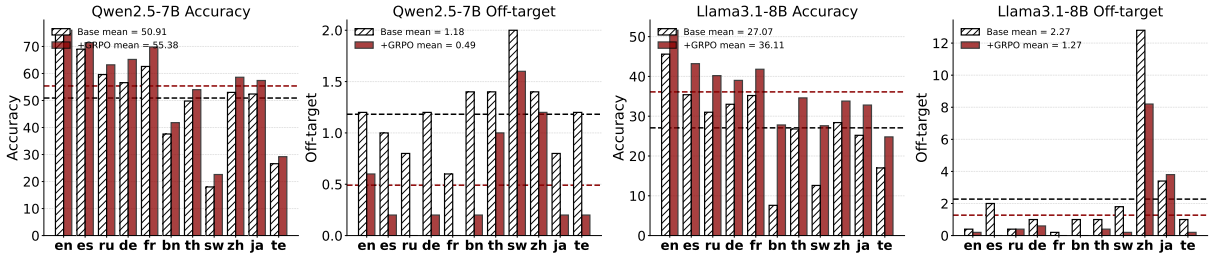


Figure 2: **Effect of model family under controlled post-training.** Multilingual reasoning performance on MATH500. *Base* denotes the initial model, and *+GRPO* the model after English-only GRPO.

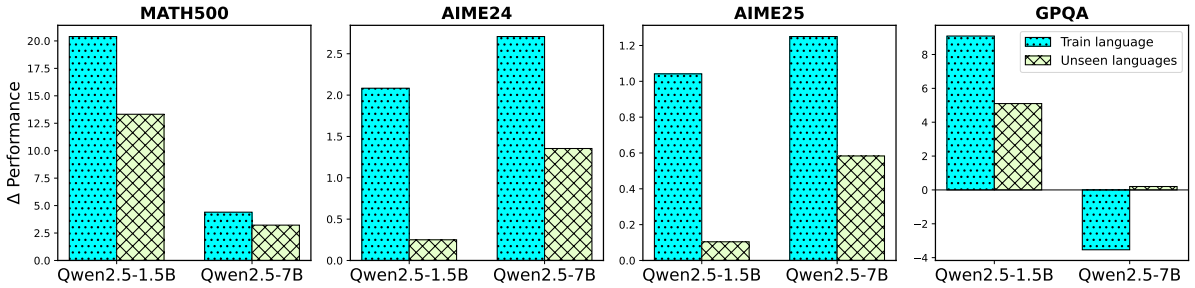


Figure 3: **Effect of model size under controlled English-only post-training.** ΔPerformance denotes the average accuracy gain over the initial model, averaged across English and unseen languages.

nation is that the small math-only training set over-specializes the 7B model toward mathematical reasoning, which can hurt out-of-domain scientific QA. Overall, the observed transfer pattern depends on the interaction between initial model strength and benchmark difficulty, rather than on model size alone.

4 Parallel Training Study

The controlled comparative results suggest that cross-lingual transfer in English-centric post-training is unstable and sensitive to initialization and alignment choices. We therefore ask ***RQ3***—*How can cross-lingual transfer be improved effectively?* In this section, we study a simple intervention: adding multilingual parallel supervision during post-training.

Our main finding is that the monolingual-to-bilingual transition yields the largest observed single-step improvement. Additional languages provide smaller but positive gains.

4.1 Setup

We use Qwen2.5-7B-Instruct as the main initial model and post-train it with GRPO on the same 1,000 English MATH examples used in Section 3, together with increasing numbers of parallel languages. The parallel problem sets are drawn from the multilingual MATH dataset described in Appendix C.1, covering seven additional languages: *es, ru, de, fr, bn, th, zh*. We validate the multilingual data with Qwen3-32B to screen for translation errors, with details and summary statistics reported in Appendix C.1. By varying the number of added languages from one to seven, we obtain a sequence of models that differ only in the amount of multilingual supervision. The full language configurations are given in Table 24.

We evaluate these models on multilingual MATH500 across eleven languages, using absolute accuracy as the primary metric and MTI as an auxiliary measure of relative transfer efficiency. Unless otherwise stated, this section focuses on multilingual MATH500, which provides the clearest view of the monolingual-to-bilingual transition in our setup.

The Power of Bilingual Training The most direct pattern in Figure 4 is that the largest observed single-step gain occurs when moving from English-only to bilingual parallel training. Using Qwen2.5-7B-Instruct as the initial model, average accuracy increases from 54.24 to 57.87, while MTI increases from 1.16 to 2.50. Importantly, this improvement is not limited to MTI. As shown in Appendix E.2, the same transition improves non-English average accuracy from 51.74 to 55.82, reduces the English–non-English gap from 27.46 to 22.58, and improves low-resource average accuracy from 29.53 to 36.27.

We refer to this observed pattern as the *First-*

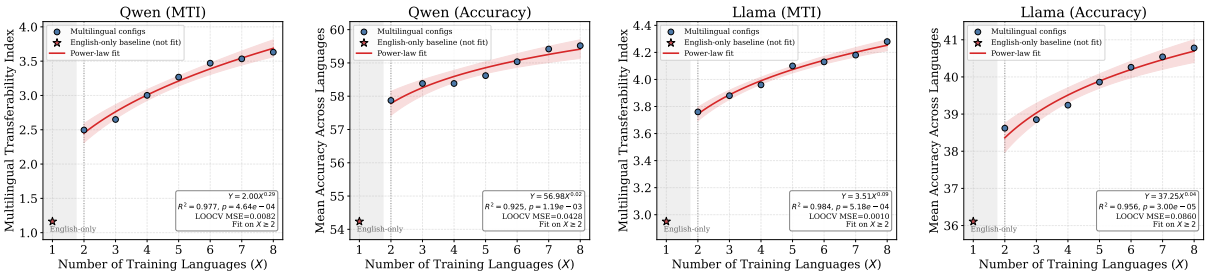


Figure 4: **Transfer improves with parallel multilingual supervision.** Each panel shows performance as a function of the number of training languages. The red star denotes the English-only baseline ($X = 1$), shown for reference but excluded from the fit. Blue points denote multilingual configurations ($X \geq 2$). The solid red curve shows a power-law-like descriptive fit over $X \geq 2$, with the shaded region indicating the 95% confidence interval. The fit summarizes the bilingual-to-multilingual trend; Appendix E.3 provides the marginal-gain analysis.

Parallel Leap. This term denotes an empirical marginal-gain pattern in our setting: the first added parallel language accounts for the largest single-step improvement, while additional languages continue to provide smaller but positive gains. Thus, moving beyond English-only supervision is especially important, while multilingual benefits do not saturate after one added language.

Beyond Bilingual: Positive but Diminishing Returns Beyond the bilingual setting, performance continues to improve as more parallel languages are added, but the gains are smaller. Figure 4 shows a smooth sub-linear trend in both MTI and average accuracy. To summarize this pattern, we fit a power-law-like function

$$f(X) = \alpha X^{\beta}, \tag{5}$$

where X is the number of training languages. We fit only configurations with $X \geq 2$ to separate the monolingual and multilingual regimes. The resulting fits are:

$$\begin{aligned} \textit{Accuracy:} \quad & f(X) = 56.98X^{0.02}, \\ \textit{Transferability:} \quad & f(X) = 2.00X^{0.29}. \end{aligned} \tag{6}$$

We use this *Parallel Scaling Trend* as a descriptive summary of the bilingual-to-multilingual regime, rather than as a universal scaling law. The sub-linear exponents indicate diminishing returns as more languages are added. To characterize the monolingual-to-bilingual transition, we rely primarily on observed marginal gains: the largest single-step improvement occurs when moving from English-only to bilingual parallel training, while additional languages continue to yield smaller but positive gains. Appendix E.3 compares alternative functional forms and reports influence diagnostics, supporting the use of the power-law-like curve as a compact descriptive fit for the multilingual regime.

4.2 Additional Analysis and Discussion

Fixed training budget. A possible alternative explanation is that parallel training helps simply because it uses more data. To test this, we fix the total number of training examples and vary only the composition. Specifically, we compare scaling English-only data (e.g., $2N$ English examples) with replacing part of it by parallel multilingual data (e.g., N English + N Russian). Here, $2N$ and $3N$ denote matched budgets of unique training examples rather than repeated epochs over the same N examples. Table 2 shows that multilingual mixtures yield stronger cross-lingual transfer than monolingual scaling alone.

Budget	Type	Acc	Off-target	MTI
1×	1×En	54.24	0.49	1.16
2×	2×En	55.82	0.58	1.92
2×	1×En + 1× Ru	57.87	0.20	2.50
3×	3×En	57.13	0.42	2.38
3×	1×En + 1× Ru + 1× Fr	58.38	0.24	2.65

Table 2: The Fixed-Budget ablation experiment based on Qwen2.5-7B-Instruct.

Parallel vs. non-parallel data. We also compare parallel bilingual supervision against non-parallel multilingual data. This comparison separates multilingual exposure from cross-lingual alignment. In the parallel setting, the model sees aligned versions of the same mathematical problems across languages, whereas in the non-parallel setting, it sees different problems in different languages and therefore has greater problem diversity. An exposure- or diversity-based explanation would predict comparable gains from non-parallel multilingual supervision. Figure 6 shows that parallel data yields stronger gains, suggesting that same-problem cross-

lingual alignment is an important part of the bilingual improvement. Additional group-level statistics are provided in Appendix E.4.

Does the chosen language matter? A practical question is whether the gain depends strongly on which bilingual language is added. Figure 7 shows that adding one parallel language consistently improves transfer, with modest differences across Russian, Bengali, German, and Chinese. Thus, the bilingual gain is robust to the choice of added language, although some language-specific effects remain.

Scaling the amount of training data We also study whether the benefit of parallel training persists as the amount of training data increases. Figure 5 compares the 1K and 3K training data for each language under English-only, bilingual, and trilingual training settings. Increasing the training data improves both MTI and average accuracy, while the relative pattern remains stable: the largest gain comes from moving from monolingual to bilingual training, and adding a second parallel language yields a smaller additional improvement.

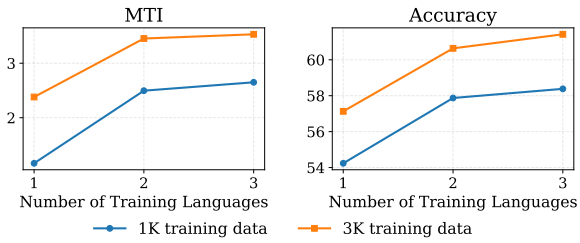


Figure 5: Scaling training data based on Qwen2.5-7B-Instruct.

Prompt ablation. A possible concern is that the bilingual gains in the main setting are driven by explicit language-control prompting. To test this, we compare two evaluation protocols: *language-controlled prompting*, used in the main experiments, and *no language control*, where all explicit target-language instructions are removed. We evaluate two Qwen2.5-7B-Instruct-based models: an *English-only* GRPO model and an *English+Ru* parallel-trained GRPO model. Table 3 shows that the bilingual gains are not reducible to test-time prompt control alone. Without language-control prompts, the English-only model exhibits severe language drift, whereas the parallel-trained model maintains near-zero off-target behavior and slightly higher accuracy. Under language-controlled evaluation, the parallel-trained model also remains stronger in accuracy.

Evaluation	Training	Acc	Off-target
No language control	English-only	57.82	75.76
No language control	English+Ru	58.43	0.36
Language-controlled	English-only	54.20	0.50
Language-controlled	English+Ru	57.90	0.20

Table 3: **Prompt ablation.** Average accuracy and off-target rate on multilingual MATH500 under two evaluation protocols.

Extension across model families and source languages. We repeat the analysis on Llama3.1-8B-Instruct and observe the same qualitative pattern in Figure 4: the largest gain comes from moving beyond English-only training. We further test a Chinese-centric variant in Table 4. The same trend appears there as well: adding one parallel language yields the largest improvement, while adding another language provides a smaller additional gain. These results suggest that the bilingual gain is not specific to a single model family or source language.

Training Data	Acc	Off-target	MTI
Zh	55.00	0.38	0.86
Zh + Ru	$58.13_{(+3.13)}$	$0.26_{(-0.12)}$	$1.69_{(+0.83)}$
Zh + Ru + Fr	$58.76_{(+3.76)}$	$0.23_{(-0.15)}$	$1.87_{(+1.01)}$

Table 4: **Chinese-centric parallel training results** on multilingual MATH500.

Robustness to transfer metric choice. We further examine whether the observed monolingual-to-bilingual gain depends on the specific formulation of MTI. To this end, we repeat the main parallel-training analysis using two alternative transfer metrics, normalized-gain ratio and raw-gain ratio in Table 14, and additionally test the sensitivity of the results to low-base-accuracy cases by excluding unseen-language cases below different base-accuracy thresholds in Table 15. While the absolute values differ across metrics and filters, the main qualitative result remains unchanged: the largest improvement still comes from moving from monolingual to bilingual training. This suggests that the bilingual gain is not specific to the original MTI definition (Appendix F.1).

5 Related Work

Reasoning Transfer Recent large reasoning models (LRMs) have been driven by reinforcement-

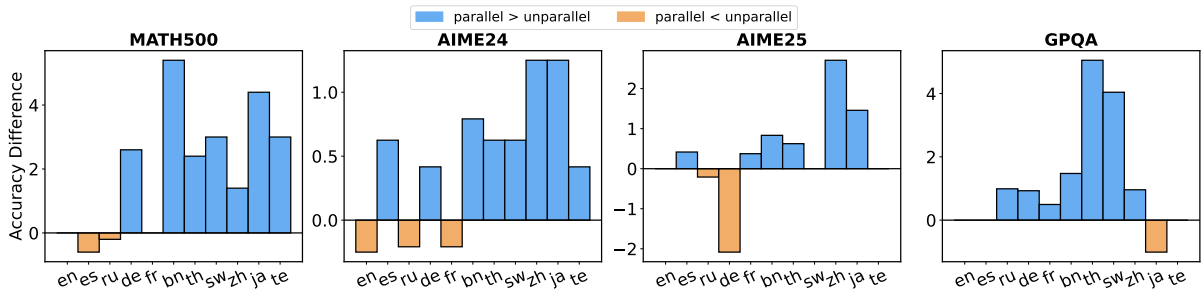


Figure 6: **Parallel vs. non-parallel multilingual supervision.** Accuracy differences under parallel and non-parallel data training based on Qwen2.5-7B-Instruct.

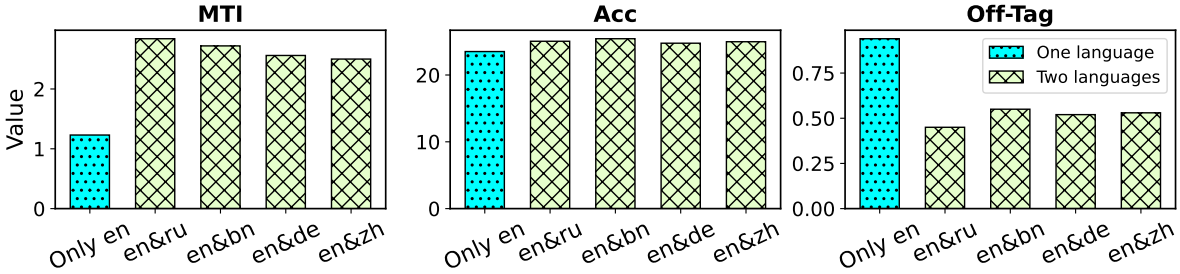


Figure 7: **The impact of the selected parallel language.** Multilingual reasoning performance across different single-language parallel training settings based on Qwen2.5-7B-Instruct.

learning-based post-training. OpenAI's O1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025) show that RL can substantially improve mathematical reasoning, especially when combined with verifiable or rule-based rewards (Shao et al., 2024). As these models improve, an important question is how well the acquired reasoning behavior transfers beyond the training setting. Prior work has mainly studied this problem across tasks (Hu et al., 2025a), domains (Huan et al., 2025; Chu et al., 2025), or modalities (Liu et al., 2025a), showing that RL-induced gains may not transfer uniformly beyond the training setting. Our work is complementary to this line of research: instead of cross-task or cross-modal transfer, we focus on cross-lingual transfer of mathematical reasoning.

Cross-Lingual Transfer Cross-lingual transfer in English-centric LLMs has been widely studied in multilingual NLP, alignment, and reasoning settings. Prior work shows that English-trained reward models (Wu et al., 2024; Hong et al., 2025), preference alignment (Yang et al., 2025b,a), and lightweight multilingual adaptation (Li et al., 2024; Chirkova and Nikoulina, 2024) can often generalize beyond English. In multilingual reasoning, recent studies examine model merging for mathematical capability transfer (Bandarkar et al., 2025), cross-lingual test-time scaling (Yong et al., 2025), response-language control (Qi et al., 2025), language–reasoning disentanglement (Zhao et al., 2025), and RL training on non-English data (Huang et al., 2025a). Parallel multilingual data has also been shown to improve multilingual capability in prompting and pretraining settings (Mu et al., 2024; Qorib et al., 2025). Our work is complementary: we study cross-lingual transfer in reasoning-oriented post-training, compare same-problem parallel supervision with non-parallel multilingual exposure, and analyze how transfer changes as the number of parallel training languages increases.

6 Conclusion

We study cross-lingual mathematical reasoning transfer in English-centric LLMs under multilingual evaluation. Our results show that transfer exists but is highly variable, and that English reasoning performance alone does not reliably predict multilingual performance. Through controlled GRPO-based experiments, we find that stronger initialization can improve final multilingual accuracy without necessarily improving relative transfer efficiency. We further show that parallel multilingual supervision improves transfer in our setting, with the largest observed single-step gain from English-only to bilingual training. Overall, our findings highlight the need to evaluate multilingual reasoning directly using both absolute performance and relative transfer-efficiency metrics.

Limitations

Our study has several limitations. First, our controlled experiments remain small-scale compared with production-level reasoning post-training, even though we extend the training data from 1K to 3K examples and observe the same qualitative pattern. The results should therefore be interpreted as evidence under our controlled setup rather than as claims about large-scale post-training dynamics. In addition, our controlled experiments are primarily based on GRPO, so they should not be read as a general comparison of all RL-style post-training methods against SFT. Second, our multilingual evaluation relies on translated benchmarks, which may introduce language-dependent artifacts such as wording shifts, notation mismatch, or uneven difficulty across languages. Evaluation is also conducted under language-controlled prompting, so the reported gains may reflect language compliance and response-format stability in addition to task-solving ability. Finally, MTI is used only as an auxiliary diagnostic of relative transfer efficiency. It can be unstable in low-accuracy regimes and should not be interpreted as a measure of absolute multilingual reasoning ability. Similarly, the power-law-like trend in the parallel-training study is intended as a descriptive summary of the observed bilingual-to-multilingual regime, rather than as a theoretical scaling law. Overall, our controlled results establish empirical associations rather than a definitive causal mechanism.

Ethical Considerations

This work studies the cross-lingual transfer of mathematical reasoning in English-centric LLMs. Our contributions are primarily empirical and methodological: we analyze transfer behavior under controlled multilingual evaluation and study training interventions that may improve multilingual reasoning performance. We use publicly available benchmarks that do not contain sensitive personal information to the best of our knowledge. At the same time, multilingual evaluation based on translated benchmarks may overstate or understate performance in some languages if translation artifacts, notation mismatch, or mixed-language mathematical responses are not carefully accounted for. Improvements in response-language control should therefore not be interpreted as complete evidence of robust multilingual reasoning in real-world settings.

References

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. 2025. Layer swapping for zero-shot cross-lingual transfer in large language models. In *The Thirteenth International Conference on Learning Representations*.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 695–708.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. In *Forty-second International Conference on Machine Learning*.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, et al. 2025. Openthoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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 *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Jiwoo Hong, Noah Lee, Rodrigo Martínez-Castaño, César Rodríguez, and James Thorne. 2025. Cross-lingual transfer of reward models in multilingual alignment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 82–94.

Chuxuan Hu, Yuxuan Zhu, Antony Kellermann, Caleb Biddulph, Suppakit Waiwitlikhit, Jason Benn, and Daniel Kang. 2025a. Breaking barriers: Do reinforcement post training gains transfer to unseen domains? *arXiv preprint arXiv:2506.19733*.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025b. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.

Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. 2025. Does math reasoning improve general llm capabilities? understand-

ing transferability of llm reasoning. *arXiv preprint arXiv:2507.00432*.

Shulin Huang, Yiran Ding, Junshu Pan, and Yue Zhang. 2025a. Beyond english-centric training: How reinforcement learning improves cross-lingual reasoning in llms. *arXiv preprint arXiv:2509.23657*.

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Team Kimi, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Chong Li, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024. X-instruction: Aligning language model in low-resource languages with self-curated cross-lingual instructions. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 546–566.

Junxiao Liu, Zhijun Wang, Yixiao Li, Zhejian Lai, Liqian Huang, Xin Huang, Xue Han, Junlan Feng, and Shujian Huang. 2026. Self-improving multilingual long reasoning via translation-reasoning integrated training. *arXiv preprint arXiv:2602.05940*.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, et al. 2025a. X-reasoner: Towards generalizable reasoning across modalities and domains. *arXiv preprint arXiv:2505.03981*.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025b. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Wenyang Luo, Wayne Xin Zhao, Jing Sha, Shijin Wang, and Ji-Rong Wen. 2025. Mmath: A multilingual benchmark for mathematical reasoning. *Preprint*.

Yongyu Mu, Peinan Feng, Zhiquan Cao, Yuzhang Wu, Bei Li, Chenglong Wang, Tong Xiao, Kai Song, Tongran Liu, Chunliang Zhang, et al. 2024. Revealing the parallel multilingual learning within large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6976–6997.

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*.

Jirui Qi, Shan Chen, Zidi Xiong, Raquel Fernández, Danielle S Bitterman, and Arianna Bisazza. 2025. When models reason in your language: Controlling thinking trace language comes at the cost of accuracy. *arXiv preprint arXiv:2505.22888*.

Muhammad Reza Qorib, Junyi Li, and Hwee Tou Ng. 2025. Just go parallel: Improving the multilingual capabilities of large language models. *arXiv preprint arXiv:2506.13044*.

Abhinav Rastogi, Albert Q Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, et al. 2025. Magistral. *arXiv preprint arXiv:2506.10910*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Mingyang Wang, Lukas Lange, Heike Adel, Yunpu Ma, Jannik Strötgen, and Hinrich Schütze. 2025a. Language mixing in reasoning language models: Patterns, impact, and internal causes. *arXiv preprint arXiv:2505.14815*.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, et al. 2025b. Polymath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*.

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

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, et al. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

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

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

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Zheng-Xin Yong, M Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H Bach, and Alham Fikri Aji. 2025. Crosslingual reasoning through test-time scaling. *arXiv preprint arXiv:2505.05408*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Xue Zhang, Yunlong Liang, Fandong Meng, Songming Zhang, Kaiyu Huang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025. Think natively: Unlocking multilingual reasoning with consistency-enhanced reinforcement learning. *arXiv preprint arXiv:2510.07300*.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, et al. 2025. When less language is more: Language-reasoning disentanglement makes llms better multilingual reasoners. *arXiv preprint arXiv:2505.15257*.

Ruochen Zhou, Minrui Xu, Shiqi Chen, Junteng Liu, Yunqi Li, Xinxin Lin, Zhengyu Chen, and Junxian He. 2025. Does learning mathematical problem-solving generalize to broader reasoning? *arXiv preprint arXiv:2507.04391*.

Appendix

A Reproducibility Statement **13**

B The Usage of AI Assistants **13**

C Evaluation Details and Setup **13**
C.1 Multilingual Reasoning Benchmarks . . . 13
C.2 Evaluation Metrics . . . 13
C.3 An Overview of Open-source LRMs . . . 14

D Implementation Details **14**
D.1 GRPO Algorithm . . . 14
D.2 Training Dataset . . . 14
D.3 Experimental Environments . . . 14
D.4 Hyperparameters . . . 15

E Additional Experiments and Analysis **15**
E.1 Matched SFT-vs.-GRPO Comparison . . . 15
E.2 Absolute-Transfer Metrics under Parallel Training . . . 16
E.3 Functional-Form Comparison and Marginal-Gain Analysis . . . 16
E.4 Parallel vs. Non-Parallel Multilingual Supervision . . . 17
E.5 Additional Analysis of Open-Source Model Transfer . . . 18
E.6 Language-Level Accuracy and Relative Variance . . . 18

F Detailed Results and Analysis **18**
F.1 Robustness of Transfer Metrics . . . 18
F.2 Observational Study . . . 19
F.2.1 Template Choice for Base Models . . . 19
F.2.2 Detailed Results for Initial Model Types . . . 19
F.2.3 Detailed Results for Open-source Models . . . 19
F.3 Controlled Comparative Study . . . 19
F.3.1 Reward Hyperparameter Sensitivity . . . 19
F.3.2 Detailed Results for Initial Model Types . . . 20
F.3.3 Detailed Results for Model Family . . . 20
F.3.4 Detailed Results for Model Size . . . 20
F.4 Parallel Training Results . . . 20
F.4.1 Parallel Training Language Settings . . . 20
F.4.2 Detailed Results of Parallel Training . . . 20
F.4.3 Effect of the Selected Parallel Language . . . 20
F.5 Parallel Training from a Chinese-Centric Perspective . . . 23

G Dataset License **23**

H Prompts Template **27**
H.1 Templates for Base Model . . . 27
H.2 Multilingual Reasoning Instruction . . . 27
H.3 Prompt hacking to force response language . . . 27
H.4 Template for R1-like Reasoning . . . 27

A Reproducibility Statement

Codes and model weights will be released after review to facilitate future research. For evaluation, we follow prior works and report averaged results over 16 sampled generations per question on data-scarce benchmarks. All evaluations are conducted with temperature set to 0.6 and top-p to 0.95, with the random seed fixed to ensure deterministic outputs across runs. Note that minor variations in inference results may still occur due to differences in hardware or the version of the inference framework.

B The Usage of AI Assistants

We declare that the AI Assistants (ChatGPT and Gemini) were only used to refine the fluency of certain sentences during the writing of this paper. Every sentence polished with the LLM was carefully reviewed and approved by the authors. The LLM was not used for any other part of this research.

C Evaluation Details and Setup

C.1 Multilingual Reasoning Benchmarks

We use the multilingual version of these four reasoning benchmarks provided in (Qi et al., 2025), which use GPT-4o (Jaech et al., 2024) to translate all questions into the other ten languages *Spanish (es), Russian (ru), German (de), French (fr), Bengali (bn), Swahili (sw), Thai (th), Japanese (ja), Chinese (zh), and Telugu (te)*, resulting in a total of eleven languages for evaluation.

MATH500 The MATH500 (Hendrycks et al., 2021) benchmark assesses the mathematical reasoning and problem-solving abilities of language models, addressing the need for more challenging evaluations as their general capabilities advance. It consists of 500 problems across five core mathematical domains: algebra, combinatorics, geometry, number theory, and precalculus. Each problem is designed to test multi-step reasoning and complex problem-solving skills, going beyond simple calculations or factual recall.

AIME24&25 The AIME24 and AIME25 datasets contain problems from the American Invitational Mathematics Examination (AIME) for 2024 and 2025, respectively. AIME is a prestigious high school mathematics competition renowned for its challenging problems, consisting of 30 questions.

GPQA-Diamond GPQA-Diamond (Rein et al., 2024) consists of 198 multiple-choice questions across biology, chemistry, and physics, with difficulty levels ranging from challenging undergraduate to postgraduate. It is the highest quality subset, which includes only questions where both experts answer correctly and the majority of non-experts answer incorrectly.

Translation Validation. To assess the quality of the translated multilingual benchmarks, we perform a full-coverage LLM-based validation following prior work (Liu et al., 2026). For each benchmark and each target language, we compare every translated question against its English source using Qwen3-32B as a judge. This validation is conducted per question rather than by sampling. The judge scores each translation along four dimensions: semantic fidelity, mathematical consistency, solvability preservation, and fluency/naturalness. Based on these scores, it assigns an overall label of *Valid*, *Minor issue*, or *Major issue*, together with a short rationale and an error type when applicable (e.g., semantic shift, omission, mathematical inconsistency, or ambiguity). Across all checked translations, 99.2% are labeled as *Valid*, 0.7% as *Minor issue*, and 0.1% as *Major issue*. We use this procedure to screen for obvious translation errors and to verify that the translated benchmarks preserve the original mathematical problems at scale.

C.2 Evaluation Metrics

We use three types of evaluation metrics: task accuracy, response-language consistency, and relative transfer efficiency.

Accuracy. Accuracy measures whether the final answer is correct. Following standard mathematical reasoning evaluation, correctness is determined from the final answer only; intermediate reasoning traces are not separately scored.

Language consistency. To measure whether a model responds in the requested language, we report the off-target rate. Following LCR-style diagnostics used in multilingual reasoning benchmarks such as PolyMath and MMATH, we use LangDetect to identify the language of each generated response. Let $\hat{l}_i$ denote the language detected from the i-th generated response, and let l_i denote the requested target language. For a set of N responses,

the language consistency rate (LCR) is defined as:

$$\text{LCR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\hat{l}_i = l_i\right], \tag{7}$$

where $\mathbb{I}[\cdot]$ is the indicator function. We define the off-target rate as the complement of LCR:

$$\text{Off-target} = 1 - \text{LCR}. \tag{8}$$

Thus, lower off-target rate indicates higher response-language consistency. In this work, we report aggregate response-level off-target rate and do not separately score the language consistency of intermediate reasoning traces and final answers.

Relative transfer efficiency. We use the Multilingual Transferability Index (MTI) as an auxiliary diagnostic of relative transfer efficiency. MTI measures how much of the training-language gain is reflected in unseen languages. Let $Acc_{b,l}^{\text{trained}}$ and $Acc_{b,l}^{\text{base}}$ denote the accuracy of a trained model and its corresponding base model on benchmark b and language l. We define the relative gain on language l as:

$$\Delta R_{b,l} = \frac{Acc_{b,l}^{\text{trained}} - Acc_{b,l}^{\text{base}}}{Acc_{b,l}^{\text{base}}}. \tag{9}$$

For a training language set $\mathcal{L}_{\text{train}}$, the average relative gain is:

$$\Delta R_{b,\mathcal{L}_{\text{train}}} = \frac{1}{|\mathcal{L}_{\text{train}}|}\sum_{l\in\mathcal{L}_{\text{train}}} \Delta R_{b,l}. \tag{10}$$

For an unseen language l_{unseen}, MTI is then defined as:

$$\text{MTI}_{l_{\text{unseen}}} = \frac{1}{|B|}\sum_{b\in B} \frac{\Delta R_{b,l_{\text{unseen}}}}{\Delta R_{b,\mathcal{L}_{\text{train}}}}. \tag{11}$$

MTI should be interpreted as a relative transfer-efficiency metric rather than a measure of absolute multilingual reasoning ability. We therefore report MTI together with absolute accuracy, non-English average accuracy, low-resource accuracy, and off-target rate.

C.3 An Overview of Open-source LRMs

Table 5 provides an overview of the various open-source LLMs evaluated in our observational study. These models, which include the DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025), OpenThinker series (Guha et al., 2025), Simple-RL-Zoo series (Zeng et al., 2025), s1 series (Muennighoff et al., 2025), and DAPO-Qwen-32B (Yu et al., 2025), range in size from 1.5B to 32B.

D Implementation Details

D.1 GRPO Algorithm

GRPO is a simplified PPO-based algorithm that significantly reduces training costs by eliminating the need for a value model. It operates by sampling G rollouts $\{o_1, ..., o_G\}$ from the current policy for a given input, calculating their cumulative rewards $R = \{R_1, ..., R_G\}$, and then using these rewards to estimate advantages $\hat{A}_{i,t}$ to guide policy updates. The optimization objective for GRPO is defined as follows:

$$L_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\mathcal{L}_{i,t}^{\text{clip}}(\theta) - \beta D_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})\right)\right] \tag{12}$$

where

$$\mathcal{L}_{i,t}^{\text{clip}}(\theta) = \min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\left(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right)$$
$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})},\ \hat{A}_{i,t} = \frac{R_i - \text{mean}(R)}{\text{std}(R)} \tag{13}$$

The clipping term with ratio ε (Schulman et al., 2015) keeps the new policy close to the old one, improving training stability.

D.2 Training Dataset

The Distribution of Parallel Questions Figure 8a shows the type and level distributions of the 1,000 English training questions sampled from the MATH dataset (Hendrycks et al., 2021). The type distribution is relatively balanced, and the number of questions increases steadily from Level 1 to Level 5.

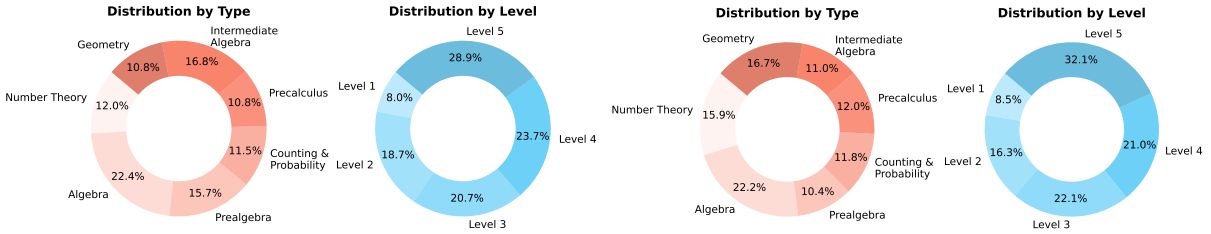
The Distribution of non-parallel Questions Moreover, Figure 8b presents the type and level distributions of 1,000 Russian questions used for a non-parallel training analysis experiment, which form a separate, non-overlapping set from the 1000 English questions. The distributions of both type and level closely match those of the English training questions. This indicates that, in the analysis comparing parallel and non-parallel training, the performance drop observed in non-parallel training is not due to distributional differences between the non-parallel and English datasets.

D.3 Experimental Environments

Training and inference were conducted on Ubuntu 22.04 with 8 × NVIDIA A800 GPUs. We use VeRL (Sheng et al., 2024) for RL training, vLLM

Model	Initial Model	Size	Training Paradigm
DeepSeek-R1-Distill-Qwen-7B	Qwen2.5-Math-7B-Base	7B	SFT
Open-Reasoner-Zero-7B	Qwen2.5-7B-Base	7B	RL
OpenThinker2-7B	Qwen2.5-7B-Instruct	7B	SFT
OpenThinker3-7B	Qwen2.5-7B-Instruct	7B	SFT
Qwen-2.5-1.5B-SimpleRL-Zoo	Qwen2.5-1.5B-Base	1.5B	RL
Qwen-2.5-7B-SimpleRL-Zoo	Qwen2.5-7B-Base	7B	RL
Qwen-2.5-14B-SimpleRL-Zoo	Qwen2.5-14B-Base	14B	RL
Qwen-2.5-Math-7B-SimpleRL-Zoo	Qwen2.5-Math-7B-Base	7B	RL
Qwen2.5-Math-7B-Dr.GRPO	Qwen2.5-Math-7B-Base	7B	RL
s1.1-7B	Qwen2.5-7B-Instruct	7B	SFT
DAPO-Qwen-32B	Qwen2.5-32B-Base	32B	RL
OpenThinker2-32B	Qwen2.5-32B-Instruct	32B	SFT
s1.1-32B	Qwen2.5-32B-Instruct	32B	SFT

Table 5: **The Overview of the Open-source LLMs Used in Observational Study**, including their initial model, parameter size, and training paradigm.



(a) (b)

Figure 8: Distribution of question difficulty. (a) The 1,000 English questions were utilized in the controlled comparative study and for parallel training. (b) The 1,000 Russian questions for non-parallel training, comprising a separate and non-overlapping set from the questions in (a).

0.8.5 (Kwon et al., 2023) for inference, and Qwen's Math evaluation codebase (Yang et al., 2024) following prior work (Zeng et al., 2025; Liu et al., 2025b).

D.4 Hyperparameters

RL Training The maximum generation length was set to 4096 tokens, and the maximum prompt length to 1024 tokens, such that their sum matches the model's maximum context length. The learning rate was fixed at 1×10^{-6}. Training was performed with a batch size of 128 questions. For each question, $G = 16$ rollouts were sampled, using a sampling temperature of 1.0. $\lambda_1 = 0.8$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.1$.

Inference In the evaluation setup, we used a temperature of 0.6, a top-p value of 0.95, and a maximum generation length of 8912 tokens for all models in the 1.5B–14B series. For 32B models, we used the same temperature (0.6) and top-p value (0.95), but set the maximum generation length to 32,768 tokens, except for DAPO-Qwen-32B, which followed the official recommended settings: a temperature of 1.0, a top-p value of 0.7, and a maximum generation length of 20,480 tokens. For AIME2024 and AIME2025, we report accuracy by averaging over 16 sampled generations per question, while for MATH500 and GPQA, accuracy is computed using a single sampled generation per question.

E Additional Experiments and Analysis

E.1 Matched SFT-vs.-GRPO Comparison

To complement the observational comparison in Section 2, we conduct a matched comparison using the same Qwen2.5-7B-Instruct backbone and the same multilingual MATH500 evaluation protocol. We compare the initial model, an SFT model trained on 1K English MATH examples, and a GRPO model trained on the same 1K English MATH examples.

Table 6 shows that GRPO achieves the strongest multilingual performance under this controlled set-

Method	En	Avg	Non-En	Low-res.	Off-target
Base	74.80	50.91	48.52	27.40	0.18
SFT	76.80	50.80	48.20	22.13	4.40
GRPO	79.20	54.24	51.74	29.53	0.49

Table 6: **Matched SFT-vs.-GRPO comparison on multilingual MATH500.** All values are percentages. All models use Qwen2.5-7B-Instruct as the backbone. SFT and GRPO are trained on the same 1K English MATH examples and evaluated under the same multilingual MATH500 protocol. Low-res. denotes the average accuracy over *bn*, *sw*, and *te*.

Setting	En	Non-En	En–Non-En Gap	Low-res.	MTI
English-only	79.20	51.74	27.46	29.53	1.163
+1 parallel language	78.40	55.82	22.58	36.27	2.496
+2 parallel languages	79.00	56.32	22.68	34.67	2.650
+3 parallel languages	77.80	56.44	21.36	34.73	3.002
+4 parallel languages	77.20	56.76	20.44	37.13	3.282
+5 parallel languages	77.40	57.18	20.22	39.33	3.475
+6 parallel languages	76.40	57.72	18.68	40.53	3.534
+7 parallel languages	76.60	57.81	18.79	40.47	3.631

Table 7: **Absolute-transfer metrics for Qwen2.5-7B-Instruct under parallel training.** All accuracy and gap values are percentages. Non-En denotes average accuracy over non-English languages. Low-res. denotes average accuracy over *bn*, *sw*, and *te*. MTI is reported as an auxiliary relative transfer-efficiency metric.

ting. Compared with SFT, GRPO improves average accuracy from 50.80 to 54.24, non-English average accuracy from 48.20 to 51.74, and low-resource average accuracy from 22.13 to 29.53. In contrast, SFT improves English accuracy over the base model but does not consistently improve multilingual performance, and it increases the average off-target rate.

This comparison supports the observational trend in Figure 1 under a matched backbone and evaluation setting. However, it is limited to one backbone, one training set, and one benchmark, and should therefore not be interpreted as a universal claim that GRPO or RL-style post-training always outperforms SFT.

E.2 Absolute-Transfer Metrics under Parallel Training

Since MTI measures relative transfer efficiency rather than absolute multilingual reasoning ability, we additionally report several absolute-transfer metrics for the Qwen2.5-7B-Instruct parallel-training experiments. Specifically, we compute non-English average accuracy, which excludes English from the language average; the English–non-English gap, which measures how far non-English performance lags behind English; and low-resource average accuracy over *bn*, *sw*, and *te*. MTI is retained only as an auxiliary diagnostic of relative transfer efficiency.

Table 7 shows that the benefit of parallel multilingual supervision is visible not only in MTI, but also in absolute multilingual performance. Moving from English-only training to one added parallel language improves non-English average accuracy from 51.74 to 55.82, reduces the English–non-English gap from 27.46 to 22.58, and improves low-resource average accuracy from 29.53 to 36.27. Adding more parallel languages further improves non-English average accuracy to 57.81 and reduces the English–non-English gap to 18.79 with seven added languages.

These results support the interpretation that MTI should be read together with absolute performance metrics. A high MTI indicates strong relative transfer efficiency, but it does not by itself imply high absolute multilingual reasoning accuracy. We therefore use MTI as a complementary diagnostic rather than as the sole measure of cross-lingual reasoning transfer.

E.3 Functional-Form Comparison and Marginal-Gain Analysis

In Section 4, we use a power-law-like curve to summarize the diminishing-return trend in the multilingual regime. This curve is intended only as a descriptive fit, not as a theoretical scaling law. To examine whether this descriptive choice is reasonable, we compare four candidate functional forms on the multilingual regime, where the number of training languages is $X = 2, \ldots, 8$:

$$\text{linear:} \quad f(X) = aX + b, \tag{14}$$

$$\text{logarithmic:} \quad f(X) = a \log X + b, \tag{15}$$

$$\text{power-law:} \quad f(X) = aX^b, \tag{16}$$

$$\text{saturation:} \quad f(X) = c - ae^{-bX}. \tag{17}$$

We evaluate the fits using R^2, leave-one-out cross-validation mean squared error (LOOCV MSE), Akaike information criterion (AIC), and Bayesian information criterion (BIC). LOOCV MSE measures out-of-sample predictive error, while AIC and BIC compare fit quality with penalties for model complexity.

As shown in Table 8, the power-law fit achieves the best LOOCV MSE, AIC, and BIC for both Qwen and Llama MTI, with competitive R^2. We therefore use the power-law-like curve as a compact descriptive summary of the multilingual

Dataset	Fit	R^2	LOOCV MSE	AIC	BIC
Qwen MTI	Linear	0.951	0.0178	-29.523	-29.631
Qwen MTI	Log	0.974	0.0121	-33.950	-34.058
Qwen MTI	Power-law	0.977	0.0082	-34.653	-34.762
Qwen MTI	Saturation	0.982	0.0894	-34.368	-34.530
Llama MTI	Linear	0.973	0.0015	-46.175	-46.284
Llama MTI	Log	0.981	0.0012	-48.701	-48.809
Llama MTI	Power-law	0.984	0.0010	-49.706	-49.814
Llama MTI	Saturation	0.000	0.0388	-18.912	-19.075

Table 8: **Curve-fitting comparison for MTI in the multilingual regime.** Fits are evaluated on configurations with $X = 2, \ldots, 8$. Lower LOOCV MSE, AIC, and BIC indicate better fit under the corresponding criterion.

Setting	$X = 1$	$X = 2$	$X = 3$	$X = 4$	$X = 5$	$X = 6$	$X = 7$	$X = 8$
Qwen MTI	2.005	0.414	0.015	0.019	0.016	0.009	0.029	0.284
Qwen Accuracy	4.503	0.512	0.099	0.001	0.007	0.009	0.008	0.072
Llama MTI	3.684	0.495	0.067	0.005	0.009	0.008	0.064	0.068
Llama Accuracy	3.341	0.862	0.004	0.015	0.000	0.000	0.005	0.029

Table 9: **Cook's distance under all-point power-law fits.** The power-law curve is fit on all observed points, $X = 1, \ldots, 8$. Larger values indicate more influential points under the fitted model.

regime, rather than as evidence for a universal scaling law.

We also examine the role of the monolingual point $X = 1$ under an all-point fit. As an influence diagnostic, we fit the power-law curve on all observed points $X = 1, \ldots, 8$ and compute Cook's distance for each point.

Table 9 shows that $X = 1$ has the largest Cook's distance in all four settings. It also exceeds the common $4/n$ rule-of-thumb threshold, which is 0.5 for $n = 8$. We treat this result as an influence diagnostic: it suggests that the monolingual point is influential under an all-point power-law fit, but it is not a formal statistical test of a separate regime.

Finally, we complement the curve-fitting analysis with a direct marginal-gain analysis. Table 10 reports the gain from each consecutive increase in the number of training languages.

Table 10 shows that the $1 \to 2$ transition yields the largest consecutive gain across average accuracy, non-English average accuracy, low-resource average accuracy, and MTI. Additional languages continue to provide gains on most metrics, but the marginal improvements are smaller.

We further summarize how much of the total $X = 1 \to 8$ gain is explained by the first added parallel language in Table 11.

Together, Tables 10 and 11 show that the English-only to bilingual transition is the largest observed consecutive improvement and explains a large fraction of the total improvement from $X = 1$ to

Transition	Avg Acc	Non-En	Low-res.	MTI
$1 \to 2$	+3.70	+4.08	+6.73	+1.333
$2 \to 3$	+0.50	+0.50	-1.60	+0.154
$3 \to 4$	+0.00	+0.12	+0.07	+0.352
$4 \to 5$	+0.20	+0.32	+2.40	+0.280
$5 \to 6$	+0.40	+0.42	+2.20	+0.193
$6 \to 7$	+0.40	+0.54	+1.20	+0.059
$7 \to 8$	+0.10	+0.09	-0.07	+0.097

Table 10: **Marginal-gain decomposition over consecutive training-language counts.** Each row reports the gain from increasing the number of training languages from X to $X + 1$ in the Qwen2.5-7B-Instruct parallel-training experiments.

Metric	$1 \to 2$	$1 \to 8$	Fraction
Avg Accuracy	+3.70	+5.30	69.8%
Non-En Avg Accuracy	+4.08	+6.07	67.2%
Low-resource Avg Accuracy	+6.73	+10.93	61.6%
MTI	+1.333	+2.468	54.0%

Table 11: **Fraction of total gain explained by the first added parallel language.** The table compares the gain from English-only to bilingual training ($1 \to 2$) with the total gain from English-only to eight-language training ($1 \to 8$).

$X = 8$: 69.8% for average accuracy, 67.2% for non-English average accuracy, 61.6% for low-resource average accuracy, and 54.0% for MTI. Additional languages continue to provide smaller but generally positive gains.

E.4 Parallel vs. Non-Parallel Multilingual Supervision

To further examine whether the bilingual gain comes from same-problem cross-lingual alignment rather than multilingual exposure alone, we compare parallel bilingual supervision with non-parallel multilingual supervision. We compute the advantage of parallel training as

$$\text{Advantage} = Acc_{\text{parallel}} - Acc_{\text{non-parallel}} \quad (18)$$

for each benchmark-language pair.

Table 12 shows that parallel supervision improves accuracy by 0.995 points on average across all benchmark-language pairs and wins on 28 out of 44 pairs. The advantage is larger for unseen non-English languages and unseen low-resource languages, while the added language itself shows only a small mean advantage. This suggests that the gain is not merely localized to the added parallel language, but transfers to other unseen languages as well.

Group	n	Mean Adv.	Wins/Ties/Losses
Overall	44	+0.995	28 / 8 / 8
Non-English	40	+1.101	28 / 5 / 7
Unseen non-English	36	+1.213	27 / 5 / 4
Unseen low-resource	12	+1.632	9 / 3 / 0
Added language (ru)	4	+0.093	1 / 0 / 3
English	4	-0.063	0 / 3 / 1

Table 12: **Parallel vs. non-parallel multilingual supervision.** The advantage is computed as $Acc_{\text{parallel}} - Acc_{\text{non-parallel}}$ over benchmark-language pairs.

E.5 Additional Analysis of Open-Source Model Transfer

We provide additional examples to clarify how Figure 1 supports the observation that cross-lingual transfer is highly variable across models, training paradigms, and target languages.

First, strong English performance does not necessarily imply strong multilingual performance. For example, on MATH500, DeepSeek-R1-Distill-Qwen-7B achieves 86.2% accuracy on English, but only 18.4% on Swahili and 17.6% on Telugu. Similarly, OpenThinker2-7B reaches 86.0% on English, but only 12.6% on Swahili and 8.4% on Telugu.

Second, transfer can differ substantially even within the same model family. For example, Qwen-2.5-7B-SimpleRL-Zoo and Qwen-2.5-14B-SimpleRL-Zoo are from the same model family, but their MATH500 average accuracies differ substantially, 56.47% versus 63.93%. Their low-resource performance also changes noticeably; for example, Telugu accuracy increases from 25.8% to 41.0%.

Third, Figure 1 suggests an observational training-paradigm trend: RL-style models have higher average MTI than SFT-style models in our open-source model pool. However, since these models differ in backbone, data, and training recipe, this pattern should be interpreted as observational rather than causal. We therefore use this analysis as motivation for the controlled comparisons in Section 3.

E.6 Language-Level Accuracy and Relative Variance

We further analyze whether languages with lower average accuracy also exhibit larger relative variation across models. For each benchmark and language, we compute the mean accuracy and coefficient of variation (CV) across open-source models:

$$\mathrm{CV} = \frac{\mathrm{std}}{\mathrm{mean}}. \tag{19}$$

Scope	Pearson r	Spearman ρ
Overall	-0.956	-0.882
MATH500	-0.971	-0.946
AIME24	-0.901	-0.946
AIME25	-0.948	-0.936
GPQA-Diamond	-0.777	-0.727

Table 13: **Correlation between mean accuracy and relative variance across languages.** For each benchmark and language, mean accuracy and CV are computed across open-source models. Negative correlations indicate that lower-accuracy languages tend to have higher relative variance.

We use CV instead of raw standard deviation to account for differences in average accuracy across languages.

Table 13 shows a strong negative correlation between language-level mean accuracy and CV across benchmarks. This indicates that lower-accuracy languages tend to exhibit higher relative variation across models. At the group level, low-resource languages (*bn*, *sw*, and *te*) also have a higher mean CV than the other languages (0.956 vs. 0.706). These results suggest that the observed target-language variation reflects not only language identity or resource level, but also higher relative instability in low-accuracy regimes.

F Detailed Results and Analysis

F.1 Robustness of Transfer Metrics

To examine whether our conclusions depend on the specific choice of MTI, we repeated the main analyses using two alternative transfer metrics.

First, we consider a normalized-gain variant:

$$\Delta N_{b,l} = \frac{Acc_{b,l}^{\text{trained}} - Acc_{b,l}^{\text{base}}}{1 - Acc_{b,l}^{\text{base}}}, \tag{20}$$

and define the corresponding transfer ratio by replacing ΔR with ΔN in the MTI computation.

Second, we consider a raw-gain ratio:

$$\Delta G_{b,l} = Acc_{b,l}^{\text{trained}} - Acc_{b,l}^{\text{base}}, \tag{21}$$

and compare the gain on an unseen language with the average gain on the training language(s):

$$\mathrm{RG}_{b,l_{\text{unseen}}} = \frac{\Delta G_{b,l_{\text{unseen}}}}{\Delta G_{b,\mathcal{L}_{\text{train}}}}. \tag{22}$$

Table 14 summarizes the corresponding results on the main parallel-training comparison. Although the absolute values differ across metrics,

the main qualitative pattern remains unchanged: moving from English-only to bilingual training yields the largest gain, and adding another parallel language provides a smaller additional improvement. This supports the main-text conclusion that the First-Parallel Leap is not specific to the original MTI definition.

Setting	MTI	NGR	RGR	Interpretation
Only English	1.163	0.396	0.732	*Baseline*
w. One parallel	2.496	0.963	1.480	*Large bilingual jump*
w. Two parallel	2.650	1.061	1.601	*Smaller additional gain*

Table 14: **Robustness of transfer metrics on the main parallel-training comparison.** We compare the original MTI with two alternative transfer metrics: normalized-gain ratio (NGR) and raw-gain ratio (RGR).

Sensitivity to base accuracy. Because MTI is based on relative gain, it can become sensitive when base accuracy is low. To assess whether the main result is driven by such cases, we recompute the transfer scores on multilingual MATH500 after excluding unseen-language cases whose base accuracy falls below different thresholds. Table 15 shows that the absolute values vary across metrics and filters, especially for the original relative-gain MTI. However, the main qualitative result remains unchanged: the transition from monolingual to bilingual training yields a clear improvement under all transfer metrics and filtering thresholds. This suggests that the bilingual gain is not an artifact of the lowest-base-accuracy cases alone.

Setting	All	$Acc^{base} > 20$	$Acc^{base} > 30$
	Relative-gain MTI		
Only English	1.16	1.16	1.23
+ One parallel	2.50	2.42	1.88
+ Two parallel	2.65	2.62	2.13
+ Three parallel	3.00	3.16	2.58
+ Four parallel	3.28	3.25	2.34
+ Five parallel	3.48	3.41	2.02
	Normalized-gain ratio		
Only English	0.40	0.40	0.48
+ One parallel	0.96	0.95	1.05
+ Two parallel	1.06	1.08	1.22
+ Three parallel	1.15	1.19	1.37
+ Four parallel	1.18	1.20	1.33
+ Five parallel	1.63	1.57	1.85
	Raw-gain ratio		
Only English	0.84	0.84	0.85
+ One parallel	1.45	1.47	1.46
+ Two parallel	1.61	1.67	1.68
+ Three parallel	1.95	2.04	1.97
+ Four parallel	1.85	1.88	1.81
+ Five parallel	2.06	2.11	2.08

Table 15: **Sensitivity of transfer scores to low-base-accuracy cases on multilingual MATH500.** We recompute the main transfer scores after excluding unseen-language cases whose base accuracy falls below different thresholds.

F.2 Observational Study

F.2.1 Template Choice for Base Models

To evaluate base models consistently, we compare three template settings for *Qwen2.5-7B-Base* and *Qwen2.5-Math-7B-Base*: `Qwen-Math Template`, `Qwen-Instruct Template`, and `No Template`. As shown in Table 16, the `Qwen-Instruct Template` yields the best overall reasoning accuracy and target-language consistency on multilingual MATH500. We therefore use it as the default template for evaluating all general-base and math-base models.

F.2.2 Detailed Results for Initial Model Types

Table 17 reports the full multilingual accuracy and off-target results for the initial models used in our analysis. Two patterns are worth noting. First, within the Qwen2.5-7B series, the instruction-tuned model has the lowest off-target rate and the highest final multilingual accuracy. Second, within the Qwen2.5-Base series, larger models generally achieve higher multilingual accuracy and lower off-target rates. These results support the main-text observation that stronger final multilingual performance and stronger relative transfer efficiency are not always aligned.

F.2.3 Detailed Results for Open-source Models

Tables 18 and 19 provide the full transfer, accuracy, and off-target results for the open-source models in the observational study. These tables complement Figure 1 in the main text by showing the benchmark-level and language-level variation behind the averaged results.

F.3 Controlled Comparative Study

F.3.1 Reward Hyperparameter Sensitivity

To examine the role of different reward components, we vary the weights on the accuracy, format, and language-consistency rewards. Table 20 reports the corresponding results. Performance is

Settings	Accuracy per language											Average	
	en	es	ru	de	fr	bn	th	sw	zh	ja	te	Acc	Off-target
Qwen2.5-7B-Base													
Qwen-Math Template	49.2	31.8	25.2	28.2	30.0	5.8	23.0	2.4	27.0	15.2	1.4	21.7	16.6
Qwen-Instruct Template	50.6	38.0	30.0	33.2	38.4	10.4	26.8	2.4	30.0	27.8	4.4	26.5	15.7
No Template	44.4	38.2	28.2	28.4	35.6	6.0	17.0	0.2	29.2	19.8	1.2	22.6	18.5
Qwen2.5-Math-7B-Base													
Qwen-Math Template	43.4	36.6	2.8	21.8	21.4	26.2	15.0	2.2	36.6	9.4	1.2	19.7	30.5
Qwen-Instruct Template	56.6	46.4	11.2	33.4	36.8	31.4	28.2	4.2	44.4	25.2	3.2	29.2	18.0
No Template	37.8	33.0	4.2	29.8	37.4	29.0	12.4	0.0	36.2	17.2	3.0	21.8	31.5

Table 16: **The Performance of Base Models with Different Template Settings.** Accuracy (%) and Off-target rate (%) across languages for different template settings on multilingual MATH500 benchmark.

most sensitive to the accuracy reward and the format reward. Removing the language-consistency reward substantially increases the off-target rate, while removing the format reward leads to the largest overall degradation. We use this analysis as a sensitivity check on the reward design; it does not fully disentangle reasoning improvement from formatting or language compliance.

F.3.2 Detailed Results for Initial Model Types

Table 21 reports the full results for the initial model type comparison in the controlled comparative study.

F.3.3 Detailed Results for Model Family

Figure 9 provides the full benchmark-level results for the model family comparison between Qwen2.5-7B-Instruct and Llama3.1-8B-Instruct. In both families, English-only GRPO improves performance across languages, but the relative gains differ substantially. As discussed in the main text, this comparison should be interpreted as a controlled comparison under matched post-training settings rather than as a pure estimate of family effects, since tokenization, pretraining data, and architecture remain entangled.

Cross-lingual Transfer within the Sino-Tibetan Family. To complement the model family comparison, we also evaluate Qwen2.5-7B-Instruct and Meta-Llama-3.1-8B-Instruct on several Sino-Tibetan languages after English-only GRPO. Table 22 shows that Qwen retains higher absolute accuracy in Chinese, while Llama exhibits larger relative gains on lower-performing languages such as Tibetan and Myanmar. We report this result as an additional case study consistent with the main-text observation that stronger starting performance and larger relative gains do not always coincide.

F.3.4 Detailed Results for Model Size

Table 23 reports the full results for the model size comparison between *Qwen2.5-1.5B-Instruct* and *Qwen2.5-7B-Instruct*. The smaller model shows larger gains on multilingual MATH500 and GPQA-Diamond, whereas the larger model attains stronger final performance on the harder AIME benchmarks. These results support the main-text observation that the effect of model scale depends on the interaction between initial model strength and benchmark difficulty.

F.4 Parallel Training Results

F.4.1 Parallel Training Language Settings

Table 24 lists the language configurations used in the parallel training study, where the number of training languages increases from 1 to 7.

F.4.2 Detailed Results of Parallel Training

Table 25 reports the full accuracy results across languages for different numbers of training languages. These tables complement Figure 4 in the main text by providing the full language-level breakdown behind the averaged trends.

F.4.3 Effect of the Selected Parallel Language

Figures 10 report benchmark-level results for different choices of the added parallel language. Across settings, the same qualitative pattern remains: adding one parallel language consistently improves transfer, and the differences across language choices are modest relative to the monolingual-to-bilingual jump. Some language-specific effects remain, especially in lower-resource languages, but the main bilingual gain is robust to the choice of the added language.

Settings	Accuracy per language											Average	
	en	es	ru	de	fr	bn	th	sw	zh	ja	te	Acc	Off-target
Multilingual MATH500													
Qwen2.5-1.5B	19.60	10.60	7.80	1.00	9.80	1.00	3.00	0.00	8.60	2.20	0.00	5.78	21.02
Qwen2.5-Math-7B	56.60	46.40	11.20	33.40	36.80	31.40	28.20	4.20	44.40	25.20	3.20	29.18	17.96
Qwen2.5-7B-Instruct	74.80	69.00	59.60	56.60	62.60	37.60	49.80	18.00	53.00	52.40	26.60	50.91	0.18
Qwen2.5-7B	50.60	38.00	30.00	33.20	38.40	10.40	26.80	2.40	30.00	27.80	4.40	26.55	15.69
Qwen2.5-14B	42.20	40.40	36.00	29.60	36.20	22.60	26.60	5.60	18.80	25.20	5.60	26.25	15.55
Qwen2.5-32B	54.00	50.80	42.00	37.80	46.80	24.60	33.00	13.60	28.00	42.60	11.60	34.98	5.56
Qwen2.5-32B-Instruct	78.60	73.60	68.00	68.40	69.40	53.20	60.60	37.40	61.40	65.00	43.40	61.73	0.13
Multilingual AIME24													
Qwen2.5-1.5B	0.21	0.42	0.00	0.00	0.00	0.00	0.21	0.00	0.63	0.00	0.00	0.13	19.26
Qwen2.5-Math-7B	13.75	6.46	1.67	3.33	4.79	2.71	3.33	0.00	6.88	1.67	0.63	4.11	21.76
Qwen2.5-7B-Instruct	10.42	8.96	8.13	8.75	8.54	2.92	4.58	1.04	5.63	4.38	1.88	5.93	0.34
Qwen2.5-7B	2.29	1.67	2.08	2.71	2.08	0.21	0.63	0.00	1.25	0.63	0.00	1.23	9.89
Qwen2.5-14B	2.50	2.50	2.29	2.50	2.71	0.63	0.21	0.00	1.04	0.83	0.21	1.40	16.02
Qwen2.5-32B	2.71	3.13	2.08	3.33	2.71	0.21	1.04	0.00	1.25	2.08	0.00	1.69	4.07
Qwen2.5-32B-Instruct	15.63	12.71	11.25	11.67	12.08	5.42	7.50	2.92	7.29	10.63	2.50	9.05	0.51
Multilingual AIME25													
Qwen2.5-1.5B	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.00	0.00	0.02	61.14
Qwen2.5-Math-7B	6.04	3.13	0.83	1.46	2.29	0.42	0.83	0.00	4.79	0.42	0.42	1.88	23.47
Qwen2.5-7B-Instruct	7.08	5.63	5.21	3.96	4.79	0.83	1.67	0.00	3.75	2.71	0.42	3.28	0.55
Qwen2.5-7B	0.83	0.83	0.21	1.25	0.63	0.21	0.00	0.00	0.42	0.21	0.00	0.42	10.38
Qwen2.5-14B	1.25	2.08	2.50	1.67	1.46	0.00	0.21	0.00	0.42	1.46	0.21	1.02	16.99
Qwen2.5-32B	1.04	1.04	1.46	0.21	0.83	0.21	0.00	0.00	1.04	1.04	0.21	0.64	4.02
Qwen2.5-32B-Instruct	11.25	7.29	7.29	6.04	6.67	1.25	2.71	0.00	5.42	2.71	0.42	4.64	0.30
Multilingual GPQA-Diamond													
Qwen2.5-1.5B	15.66	14.65	15.15	16.67	11.62	7.58	15.15	13.64	15.15	6.06	15.15	13.31	20.02
Qwen2.5-Math-7B	16.16	13.64	3.54	17.17	15.66	17.68	17.68	11.62	22.22	0.51	16.16	13.82	27.18
Qwen2.5-7B-Instruct	36.36	32.83	23.74	36.36	33.84	26.77	29.80	24.75	30.81	29.80	21.72	29.71	0.64
Qwen2.5-7B	28.79	24.24	20.71	22.22	18.18	11.11	21.72	17.68	23.23	17.17	12.63	19.79	9.69
Qwen2.5-14B	26.26	15.15	20.71	21.21	27.78	20.20	24.24	22.22	12.12	10.61	16.16	19.70	20.98
Qwen2.5-32B	28.28	30.30	28.28	33.33	23.74	15.66	24.75	17.68	27.27	27.78	17.17	24.93	9.00
Qwen2.5-32B-Instruct	45.45	41.92	38.89	41.41	44.44	29.80	38.38	32.83	36.36	38.38	27.27	37.74	0.28

Table 17: **The Performance of Initial Models.** Accuracy (%) and Off-target rate (%) across languages for different Initial models.

Models	Multilingual Reasoning Benchmarks				MTI		
	MATH500	*AIME24*	*AIME25*	*GPQA-D*	*ID*	*OOD*	*Avg*
DeepSeek-R1-Distill-Qwen-7B	3.493	2.312	2.864	4.168	3.493	3.115	3.209
Open-Reasoner-Zero-7B	3.195	2.677	1.479	1.320	3.195	1.825	2.168
OpenThinker2-7B	0.093	0.876	1.843	1.604	0.093	1.441	1.104
OpenThinker3-7B	0.157	1.502	2.434	1.318	0.157	1.752	1.353
Qwen-2.5-1.5B-SimpleRL-Zoo	5.322	2.173	0.856	1.383	5.322	1.383	3.353
Qwen-2.5-7B-SimpleRL-Zoo	4.543	3.189	1.217	6.531	4.543	3.646	3.870
Qwen-2.5-14B-SimpleRL-Zoo	2.381	3.360	0.959	1.655	2.381	1.991	2.089
Qwen-2.5-Math-7B-SimpleRL-Zoo	3.920	2.884	3.079	4.335	3.920	3.433	3.555
Qwen2.5-Math-7B-Dr.GRPO	2.807	1.324	3.158	2.149	2.807	2.210	2.359
s1.1-7B	0.310	0.920	1.192	0.671	0.310	0.928	0.773
DAPO-Qwen-32B	3.634	2.337	2.066	0.854	3.634	1.752	2.223
OpenThinker2-32B	0.936	1.513	4.235	0.201	0.936	1.983	1.721
S1.1-32B	1.382	1.583	3.429	0.821	1.382	1.944	1.804

Table 18: **The Performance of Various Open-source Models. Part 1:** Multilingual Transferability Index (MTI) of various models across benchmarks. The columns ID, OOD, and Avg refer to the MTI on in-domain (MATH500), out-of-domain (AIME24, AIME25, GPQA-Diamond), and all tasks, respectively.

Settings	Accuracy per language											Average	
	en	es	ru	de	fr	bn	th	sw	zh	ja	te	Acc	Off-target
Multilingual MATH500													
DeepSeek-R1-Distill-Qwen-7B	86.20	76.20	67.00	66.40	69.80	40.20	53.80	18.40	70.00	53.20	17.60	56.25	2.69
Open-Reasoner-Zero-7B	81.60	76.00	70.40	69.80	71.20	45.20	63.20	15.00	62.80	61.80	17.60	57.69	5.20
OpenThinker2-7B	86.00	74.80	64.80	63.00	73.00	34.40	58.80	12.60	65.80	70.00	8.40	55.60	36.13
OpenThinker3-7B	85.80	81.80	75.00	69.20	76.40	17.00	48.80	17.40	69.60	65.00	10.40	56.04	60.75
Qwen-2.5-1.5B-SimpleRL-Zoo	57.60	41.40	36.80	38.20	42.40	11.80	28.80	14.60	33.60	31.00	7.40	31.24	21.51
Qwen-2.5-7B-SimpleRL-Zoo	77.60	72.00	65.00	64.20	68.00	42.20	60.40	24.20	62.00	59.80	25.80	56.47	4.31
Qwen-2.5-14B-SimpleRL-Zoo	82.40	74.40	71.00	69.40	73.60	54.80	69.00	33.20	65.60	68.80	41.00	63.93	0.47
Qwen-2.5-Math-7B-SimpleRL-Zoo	80.40	72.60	66.00	68.60	70.60	45.80	57.40	15.00	61.80	54.40	14.20	55.16	8.33
Qwen2.5-Math-7B-Oat-Zero	79.80	72.40	32.80	55.60	50.40	47.60	49.40	16.80	58.00	43.40	11.80	47.09	15.11
s1.1-7B	75.80	68.20	60.80	59.00	69.60	37.00	57.40	16.40	57.20	51.60	20.40	52.13	12.76
DAPO-Qwen-32B	68.80	65.00	58.80	60.20	63.00	52.80	58.40	44.80	54.20	56.80	44.80	57.05	11.85
OpenThinker2-32B	96.00	88.60	85.20	84.20	85.00	73.00	80.60	35.40	77.40	75.60	47.20	75.29	13.02
S1.1-32B	95.40	91.20	85.00	83.60	88.00	73.00	81.20	57.00	77.40	80.80	53.60	78.75	27.40
Multilingual AIME24													
DeepSeek-R1-Distill-Qwen-7B	40.63	27.71	26.25	23.13	27.50	6.46	9.79	2.50	30.42	9.79	0.83	18.64	7.77
Open-Reasoner-Zero-7B	16.25	18.13	17.29	15.21	17.71	9.58	14.79	1.67	14.79	14.79	1.25	12.86	10.78
OpenThinker2-7B	37.08	18.33	17.08	13.75	20.83	11.04	20.21	2.50	25.63	27.29	5.42	18.11	39.72
OpenThinker3-7B	26.25	32.08	23.54	26.46	29.38	5.63	16.25	3.54	26.46	19.38	3.54	19.32	63.28
Qwen-2.5-1.5B-SimpleRL-Zoo	0.00	0.00	0.42	0.00	0.21	0.00	0.21	0.42	1.25	0.21	0.00	0.25	60.42
Qwen-2.5-7B-SimpleRL-Zoo	6.25	7.29	6.25	7.29	8.33	3.75	5.42	2.08	5.42	4.38	2.50	5.36	77.16
Qwen-2.5-14B-SimpleRL-Zoo	12.71	13.13	13.13	10.42	13.33	9.17	9.79	3.54	10.42	11.46	5.63	10.25	0.42
Qwen-2.5-Math-7B-SimpleRL-Zoo	25.83	15.42	13.96	11.25	13.33	5.00	8.13	1.67	10.21	8.54	2.50	10.53	11.29
Qwen2.5-Math-7B-Oat-Zero	28.33	12.92	5.42	8.54	11.67	6.25	8.75	1.04	12.29	6.67	0.42	9.30	20.57
s1.1-7B	14.38	10.42	10.42	10.21	11.46	5.21	7.92	1.67	8.75	7.71	0.21	8.03	7.95
DAPO-Qwen-32B	54.58	50.00	51.67	46.04	50.00	42.50	36.04	19.17	40.83	45.42	27.29	42.14	5.91
OpenThinker2-32B	74.17	61.88	55.42	56.67	55.42	59.17	49.17	13.96	56.88	37.71	34.58	50.45	22.08
S1.1-32B	58.75	55.21	49.17	51.25	53.33	36.04	41.25	19.38	44.58	46.88	17.08	42.99	7.80
Multilingual AIME25													
DeepSeek-R1-Distill-Qwen-7B	29.58	20.21	21.25	22.29	19.79	5.63	8.75	0.42	26.67	10.00	0.00	14.96	6.97
Open-Reasoner-Zero-7B	14.58	13.33	11.88	9.58	11.04	1.67	9.38	0.00	10.21	9.79	0.21	8.33	10.04
OpenThinker2-7B	28.33	21.67	20.63	17.08	21.46	9.38	17.08	2.71	25.00	24.38	2.08	17.25	39.77
OpenThinker3-7B	22.50	27.71	23.33	20.00	27.50	6.67	14.79	3.13	28.54	21.67	1.67	17.95	63.28
Qwen-2.5-1.5B-SimpleRL-Zoo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.00	0.00	0.02	61.14
Qwen-2.5-7B-SimpleRL-Zoo	4.58	3.33	2.08	3.96	5.42	0.83	3.13	0.63	1.46	2.50	0.21	2.56	79.51
Qwen-2.5-14B-SimpleRL-Zoo	13.96	11.67	10.42	10.83	10.00	3.54	6.46	1.88	6.67	8.54	2.08	7.82	0.61
Qwen-2.5-Math-7B-SimpleRL-Zoo	13.75	9.58	6.04	5.42	9.79	2.71	5.63	1.04	6.25	3.75	1.04	5.91	12.12
Qwen2.5-Math-7B-Oat-Zero	10.00	9.38	2.29	6.67	4.38	1.46	2.08	1.04	6.67	2.92	0.42	4.30	22.12
s1.1-7B	13.96	11.67	9.58	6.88	11.88	2.08	7.08	0.21	9.79	5.21	0.00	7.12	6.17
DAPO-Qwen-32B	38.13	38.54	37.29	36.25	34.58	30.83	32.71	18.33	31.67	34.17	22.29	32.25	4.56
OpenThinker2-32B	57.29	50.00	48.13	52.29	43.96	45.42	41.88	12.50	52.50	36.04	25.63	42.33	22.65
S1.1-32B	50.00	43.54	38.33	43.33	42.71	29.38	31.88	16.04	38.75	35.42	14.58	34.91	8.05
Multilingual GPQA-Diamond													
DeepSeek-R1-Distill-Qwen-7B	32.32	33.33	33.33	35.35	35.35	18.18	21.21	23.74	29.80	14.65	14.14	26.49	6.11
Open-Reasoner-Zero-7B	37.37	26.77	31.82	33.33	33.33	24.24	32.83	12.63	33.33	26.77	7.07	27.23	6.20
OpenThinker2-7B	28.79	17.68	17.17	16.67	22.73	22.22	25.76	14.14	22.22	18.69	14.14	20.02	38.15
OpenThinker3-7B	23.23	18.69	22.22	16.67	24.24	5.05	14.65	12.63	21.72	10.10	7.07	16.02	59.23
Qwen-2.5-1.5B-SimpleRL-Zoo	20.71	15.66	23.74	16.67	24.75	10.10	18.18	13.13	17.17	21.21	8.59	17.26	14.69
Qwen-2.5-7B-SimpleRL-Zoo	30.30	31.82	29.80	31.82	33.84	20.71	23.74	17.17	29.80	23.74	10.10	25.71	3.49
Qwen-2.5-14B-SimpleRL-Zoo	41.92	40.40	34.85	40.91	39.39	27.78	34.85	29.80	39.39	33.33	26.26	35.35	2.62
Qwen-2.5-Math-7B-SimpleRL-Zoo	30.81	26.26	22.73	27.78	28.28	18.18	17.17	9.09	27.27	16.67	8.08	21.12	12.26
Qwen2.5-Math-7B-Oat-Zero	25.76	17.17	7.07	21.21	15.66	19.19	21.72	9.09	30.81	6.06	12.63	16.94	19.74
s1.1-7B	17.68	14.14	20.20	22.22	29.29	9.09	17.17	16.16	24.75	16.67	18.69	18.73	11.98
DAPO-Qwen-32B	52.50	44.44	40.91	48.99	41.92	37.37	42.93	31.82	46.97	47.98	30.81	42.42	5.88
OpenThinker2-32B	62.63	57.58	58.08	59.09	58.59	50.51	47.47	21.72	56.57	0.00	0.00	42.93	22.91
S1.1-32B	64.65	57.58	57.58	59.60	56.57	41.41	48.48	36.36	56.57	53.03	32.83	51.33	11.85

Table 19: **The Performance of Various Open-source Models. Part 3:** Accuracy (%) and Off-target rate (%) across languages for various open-source models.

ID	Reward Weights			Purpose	Accuracy (Acc) per Language											Avg. Acc	Avg. Off-target	MTI
	λ_1	λ_2	λ_3		en	es	ru	de	fr	bn	th	sw	zh	ja	te			
C1	0.8	0.1	0.1	Balanced	78.4	73.4	66.0	65.6	67.2	48.8	57.4	26.2	57.8	62.0	33.8	**57.87**	**0.20**	**2.50**
C2	0.9	0.1	0.0	w/o LCR	77.8	71.4	66.2	63.4	68.0	44.2	56.4	26.2	59.6	57.8	31.8	56.62	1.24	2.10
C3	0.7	0.1	0.2	Strong LCR	77.2	70.2	65.8	65.8	67.2	43.6	55.4	25.6	60.2	56.2	32.6	56.35	0.38	2.25
C4	0.9	0.0	0.1	w/o FR	75.2	60.8	62.2	56.0	60.2	36.2	43.8	18.0	54.4	31.8	24.8	47.58	0.31	-3.45
C5	0.7	0.2	0.1	Strong FR	78.2	71.6	65.2	62.6	67.8	45.0	60.0	24.8	60.0	60.8	29.2	56.84	0.35	2.23

Table 20: Sensitivity analysis of reward hyperparameters λ, trained on English (En) with parallel Russian (Ru) data. **Acc**, **Off**, and **MTI** denote Accuracy, Off-target rate, and Multilingual Transfer Index, respectively.

Model	Average accuracy across all languages				Avg	Off-target	MTI
	MATH500	*AIME24*	*AIME25*	*GPQA*			
Qwen2.5-7B-Base	26.55	1.23	0.42	19.79	12.00	11.41	-
↬ *GRPO on En Data*	52.16	7.10	3.35	27.18	22.45	3.12	1.95
Qwen2.5-7B-Instruct	50.91	5.93	3.28	29.71	22.45	1.43	-
↬ *GRPO on En Data*	54.24	7.41	3.92	28.47	23.51	0.94	1.23
Qwen2.5-Math-7B	29.18	4.11	1.88	13.82	12.25	22.59	-
↬ *GRPO on En Data*	45.73	8.84	3.96	18.96	19.37	9.50	2.12

Table 21: **The Impact of Initial Model Type on Interventional Study.** Accuracy (%), Off-target rate (%) and MTI across different initial model types.

Model	Tibetan (bo)	Myanmar (my)	S. Chinese (zh)
Qwen2.5-7B-Instruct	10.40	9.20	53.00
+ GRPO (En)	13.60 (+3.20)	11.60 (+2.40)	58.60 (+5.60)
Llama-3.1-8B-Instruct	0.80	2.40	28.40
+ GRPO (En)	15.00 **(+14.20)**	8.40 **(+6.00)**	33.80 **(+5.40)**

Table 22: Cross-lingual transfer performance (Accuracy %) on Sino-Tibetan languages before and after English-only RPT (GRPO).

F.5 Parallel Training from a Chinese-Centric Perspective

Table 26 reports the full results for three settings: monolingual Zh training, bilingual Zh+Ru training, and trilingual Zh+Ru+Fr training. A similar qualitative pattern appears: adding one parallel language yields the largest improvement in average accuracy and MTI, while adding a second language provides a smaller additional gain. We report this result as supporting evidence that the bilingual gain is not unique to English-only post-training.

G Dataset License

We use publicly available models and benchmarks, including the multilingual versions of MATH500 (Hendrycks et al., 2021), AIME2024, AIME2025 from XReasoning (Qi et al., 2025), and GPQA-Diamond (Rein et al., 2024) from BenchMAX (Huang et al., 2025b), and follow their corresponding usage guidelines.

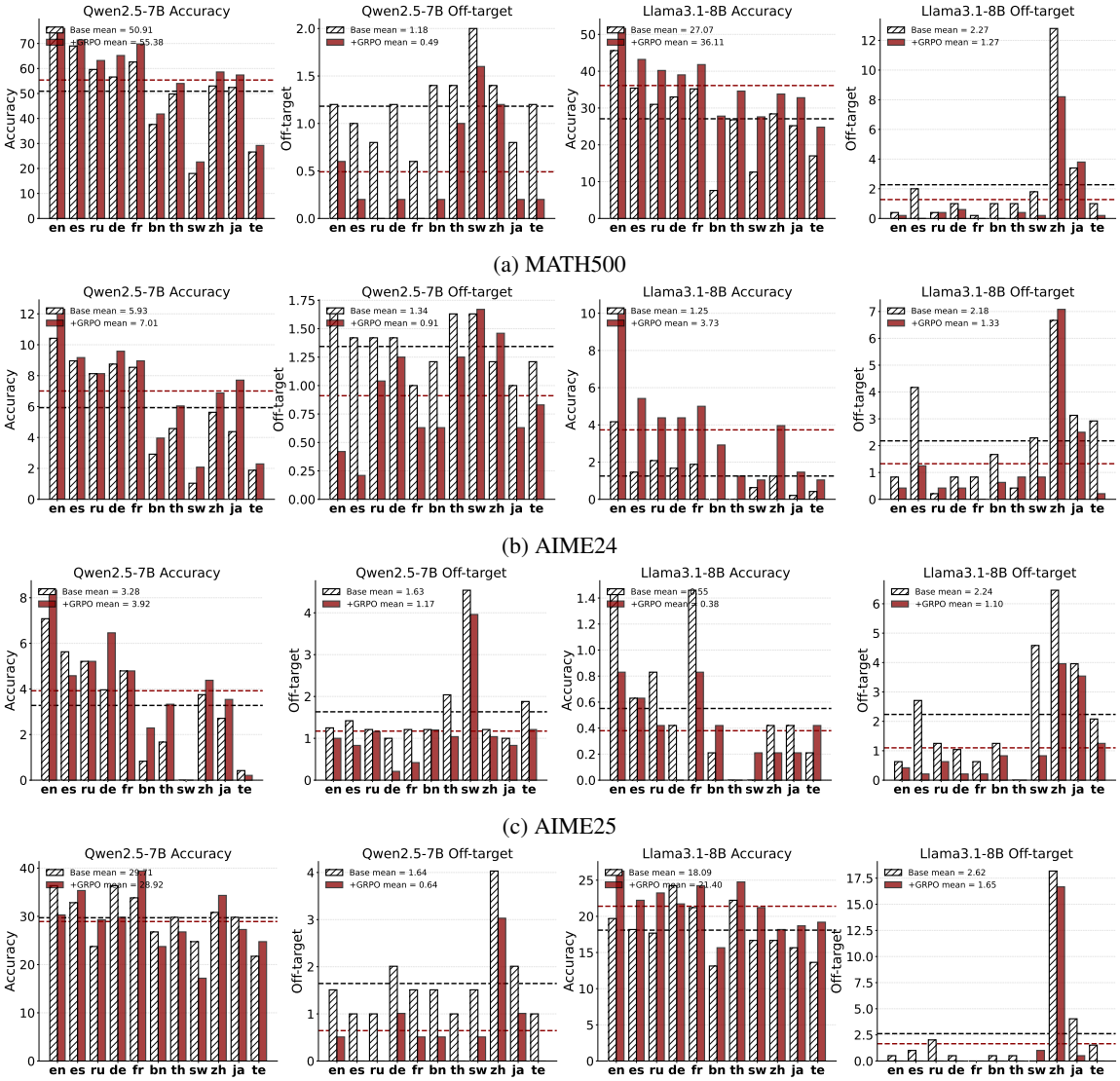


Figure 9: **The Impact of Different Model Families in Controlled Comparative Study.** Multilingual reasoning performance across languages, comparing the influence of model family using Qwen2.5-7B-Instruct and Llama3.1-8B-Instruct as initial models.

Settings	**Δ Performance**											**Average across languages**	
	en	es	ru	de	fr	bn	th	sw	zh	ja	te	*Training*	*Untraining*
Qwen2.5-1.5B-Instruct with GRPO on En Data													
MATH500	20.40	27.20	17.40	14.40	17.40	6.20	7.80	4.40	16.80	14.20	7.40	20.40	13.32
AIME24	2.08	0.42	-0.21	1.25	0.42	0.21	0.21	0.00	-0.21	0.21	0.21	2.08	0.25
AIME25	1.04	0.21	0.00	0.21	0.00	0.00	0.00	-0.21	0.42	0.21	0.21	1.04	0.10
GPQA-Diamond	9.09	20.71	-0.51	5.05	12.12	1.52	-2.53	2.02	8.59	3.54	0.51	9.09	5.10
Qwen2.5-7B-Instruct with GRPO on En Data													
MATH500	4.40	1.00	1.40	5.60	5.60	4.20	5.60	1.40	0.40	6.20	0.80	4.40	3.22
AIME24	2.71	1.04	0.83	1.46	0.42	0.42	0.42	2.08	2.08	4.58	0.21	2.71	1.35
AIME25	1.25	-1.04	0.00	2.50	0.00	1.46	1.67	0.00	0.63	0.83	-0.21	1.25	0.58
GPQA-Diamond	-3.54	4.55	0.00	-1.01	7.07	-2.02	-1.01	-7.07	4.04	-3.03	0.51	-3.54	0.20

Table 23: **The Impact of Model Size in Interventional Study.** Δ Performance on various benchmarks across *Qwen2.5-1.5B-Instruct* and *Qwen2.5-7B-Instruct*.

Settings	Training Parallel Languages
Only English	*en*
w. One parallel	*en, ru*
w. Two parallel	*en, ru, fr*
w. Three parallel	*en, ru, fr, es*
w. Four parallel	*en, ru, fr, es, de*
w. Five parallel	*en, ru, fr, es, de, bn*
w. Six parallel	*en, ru, fr, es, de, bn, th*
w. Seven parallel	*en, ru, fr, es, de, bn, th, zh*

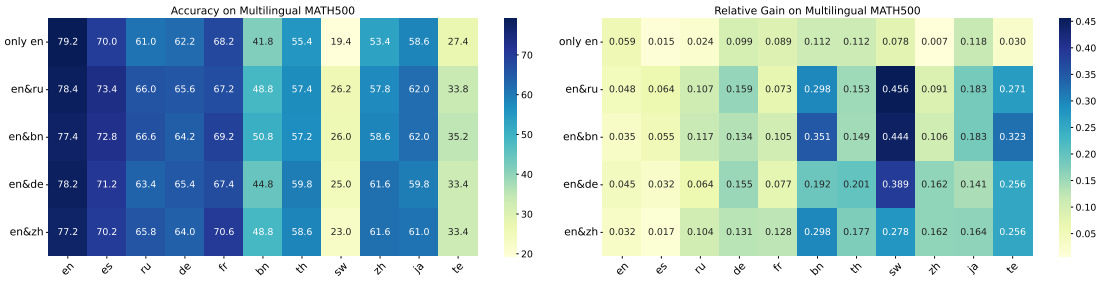
Table 24: **The Language Settings in Parallel Scaling Law.**

Settings	**Accuracy per language**											**Average**		
	en	es	ru	de	fr	bn	th	sw	zh	ja	te	**Acc**	**Off-target**	**MTI**
Multilingual MATH500														
Only English	79.2	70.0	61.0	62.2	68.2	41.8	55.4	19.4	53.4	58.6	27.4	54.2	0.5	1.163
w. One parallel	78.4	73.4	66.0	65.6	67.2	48.8	57.4	26.2	57.8	62.0	33.8	57.9	0.2	2.496
w. Two parallel	79.0	73.4	64.4	67.4	69.2	45.2	60.2	26.2	63.0	61.6	32.6	58.4	0.2	2.650
w. Three parallel	77.8	73.6	64.6	68.4	69.8	46.0	60.8	24.0	62.2	60.8	34.2	58.4	0.4	3.002
w. Four parallel	77.2	71.2	66.2	66.8	68.0	47.6	61.8	28.6	62.0	60.2	35.2	58.6	0.6	3.282
w. Five parallel	77.4	71.4	62.2	66.2	66.0	48.6	62.0	32.4	62.2	63.8	37.0	59.0	0.4	3.475
w. Six parallel	76.4	70.8	63.8	65.6	66.8	48.6	61.8	34.6	63.4	63.4	38.4	59.4	0.5	3.534
w. Seven parallel	76.6	71.2	63.6	66.2	66.2	49.4	62.6	33.8	63.5	63.4	38.2	59.5	0.2	3.631

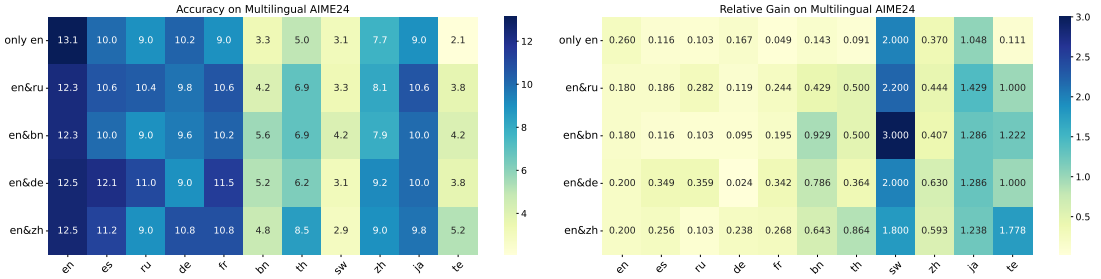
Table 25: **The Detailed Results in Parallel Scaling Law.**

Training Data	**en**	**es**	**ru**	**de**	**fr**	**bn**	**th**	**sw**	**zh**	**ja**	**te**	**Avg. Acc**	**Avg. Off-target**	**MTI**
Qwen2.5-7B-Instruct	74.8	69.0	59.6	56.6	62.6	37.6	49.8	18.0	53.0	52.4	26.6	50.91	1.18	-
Zh	76.8	70.4	61.8	62.2	66.8	43.0	55.2	21.8	59.4	56.6	31.0	55.00	0.38	0.86
Zh + Ru	78.8	72.0	64.4	65.0	70.0	44.4	60.0	28.6	61.4	60.4	33.4	58.13	0.26	1.69
Zh + Ru + Fr	78.0	72.2	63.6	66.4	69.8	45.4	61.0	30.0	62.8	64.0	33.2	**58.76**	**0.23**	**1.87**

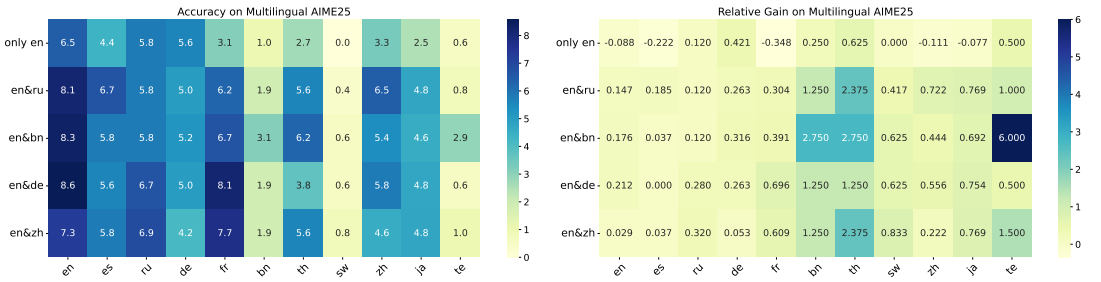
Table 26: Parallel training experiments with Chinese (Zh) as the source language. Chinese-centric parallel training results on multilingual MATH500. A similar qualitative pattern appears: adding one parallel language yields the largest gain, while adding a second language provides a smaller additional improvement.



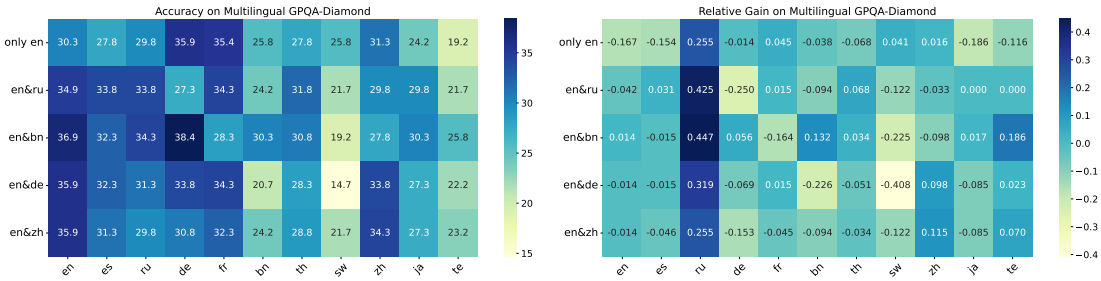
(a) MATH500



(b) AIME24



(c) AIME25



(d) GPQA-Diamond

Figure 10: **Analysis across Selected Parallel Languages.** The accuracy and relative gain across various benchmarks with different parallel languages. "Only en" denotes only fine-tuned on English data. "en&LANGUAGE" indicates the model was fine-tuned on English and a parallel language, with LANGUAGE representing *ru, bn, de, zh*, respectively.

H Prompts Template

H.1 Templates for Base Model

Qwen-Instruct Template:

```
<|im_start|>system\n
You are Qwen, created by Alibaba Cloud. You are a helpful assistant.
<|im_end|>\n
<|im_start|>user\n{instruction}<|im_end|>\n
<|im_start|>assistant\n
```

Qwen-Math Template:

```
<|im_start|>system\n
Please reason step by step, and put your final answer within \boxed{}.
<|im_end|>\n
<|im_start|>user\n{instruction}<|im_end|>\n
<|im_start|>assistant\n
```

No Template:

```
{instruction}
```

H.2 Multilingual Reasoning Instruction

The Instruction Used in Multilingual Reasoning Prompt

Please always think in `[LANGUAGE]`.

Solve the following mathematics problem step by step. At the end, provide your final answer enclosed in `\boxed{}`.

Problem: {}

H.3 Prompt hacking to force response language

The Prefixes Used in Prompt Hacking. Note that we list seven out of eleven languages.

- **English:** By request, I will start thinking in English.
- **Japanese:** 要求があれば、日本語で考え始めます。
- **Chinese:** 应要求，我将开始用中文思考。
- **Spanish:** A petición, empezaré a pensar en español.
- **French:** Sur demande, je commencerai à penser en français.
- **German:** Auf Anfrage werde ich anfangen, in Deutsch zu denken.
- **Swahili:** Kwa ombi, nitaanza kufikiria kwa Kiswahili.

H.4 Template for R1-like Reasoning

The Template for R1-like Reasoning

You are a helpful AI Assistant that provides well-reasoned and detailed responses. You first think about the reasoning process as an internal monologue and then provide the user with the answer. The final answer must be put in `\boxed{}`. Respond in the following format: `<think>\n...\n</think>\n<answer>\n...\n</answer>`