# *Best-of-L*: Cross-Lingual Reward Modeling for Mathematical Reasoning

**Sara Rajaee**[1*], **Rochelle Choenni**[2], **Ekaterina Shutova**[2], **Christof Monz**[1]
Language Technology Lab, University of Amsterdam[1]
ILLC, University of Amsterdam[2]

## Abstract

While the reasoning abilities of large language models (LLMs) continue to advance, it remains underexplored how such abilities vary across languages in multilingual LLMs and whether different languages generate distinct reasoning paths. In this work, we show that reasoning traces generated in different languages often provide complementary signals for mathematical reasoning. We propose cross-lingual outcome reward modeling, a framework that ranks candidate reasoning traces across languages rather than within a single language. Our experiments on the MGSM benchmark show that cross-lingual reward modeling improves accuracy by up to 10 points compared to using reward modeling within a single language, benefiting both high- and low-resource languages. Notably, cross-lingual sampling improves English performance under low inference budgets, despite English being the strongest individual language. Our findings reveal new opportunities to improve multilingual reasoning by leveraging the complementary strengths of diverse languages.

## 1 Introduction

Recently, many studies have focused on improving reasoning ability (Gemini Team, 2025; Hwang et al., 2025; Ranaldi and Freitas, 2024; Byun et al., 2024) or identifying key factors behind it (Ko et al., 2024; Li et al., 2025a; Liu et al., 2025). Yet, reasoning research has largely centered on English models, with multilingual models receiving comparatively little attention. Among the few, Shi et al. (2023) have shown that multilingual large language models (LLMs) have strong reasoning capabilities, even for underrepresented languages. Further progress has been made to improve multilingual math reasoning through self-consistency (Lai et al., 2025), multilingual instruction-tuning (Chen et al., 2024; Lai and Nissim, 2024), and preference optimization methods (She et al., 2024; Dang et al., 2024a; Yang et al., 2025). Building on prior work using reward modeling to enhance math reasoning in English LLMs (Cobbe et al., 2021; Shen et al., 2021; Hosseini et al., 2024; Zhang et al., 2024; Setlur et al., 2025), Hong et al. (2025) examined the transferability of English reward models to other languages. More recently, Wang et al. (2025) extended reward models to multilingual setups (multilingual reward model), but still limited sampling and scoring to a single language. We expand on this perspective by exploring how multilingual LLMs can utilize the reasoning traces generated in multiple languages for the same question, as illustrated in Figure 1.

To this end, we first study to what extent languages could potentially complement each other's mathematical reasoning skills. Interestingly, we find that even low-resource languages sometimes succeed on problems where high-resource languages fail, suggesting that their reasoning signals could provide valuable complementary information (Figure 5).

Motivated by the above finding, we propose a cross-lingual outcome reward modeling (ORM) framework to harness the *Best-of-Languages* performance, in which we train a multilingual verifier to score the correctness of reasoning traces. At inference, we sample candidate solutions in multiple languages and select the highest-scoring one. To the best of our knowledge, we are the first to propose a cross-lingual reward model that leverages complementary reasoning skills across languages.

Our results show cross-lingual ORM improves performance by over 10% and 15% compared to average results of multilingual ORM and the self-consistency baseline, respectively. Further analysis reveals that expanding the language pool consistently enhances the framework's performance.

Moreover, through an ablation study, we find that

[*]Corresponding author: s.rajaee@uva.nl



*Findings of the Association for Computational Linguistics: EACL 2026*, pages 1930–1939
March 24-29, 2026 

cross-language sampling even benefits English, especially under low-budget settings. We show that, while having English in the language pool of the cross-lingual ORM sampling positively affects the performance, some selection of non-English pools outperforms other pools containing English, supporting our argument that languages have complementary reasoning skills in multilingual language models.

## 2 Related Work

Reward modeling has become a powerful approach to enhance multilingual reasoning (Anugraha et al., 2025; Hwang et al., 2025). Prior work studied its cross-lingual transferability (Hong et al., 2025; Wu et al., 2024), where a reward model trained in one language is applied to another. These methods generally involve two stages: training the reward model (monolingual or multilingual) and sampling candidate solutions, which typically operate within a single language. For instance, *monolingual ORM* trains and samples in one language (Wang et al., 2025), while *multilingual ORM* trains the reward model on multiple languages but still ranks samples within each language. We instead introduce *cross-lingual ORM* which, like Wang et al. (2025),trains a reward model on multiple languages, but ranks solutions *across* languages for each question. Our framework is based on our observation that utilizing multiple languages has the potential to enhance reasoning performance, similar to the concurrent work (Hong et al., 2025).

## 3 Methodology

Reward models are extensively being used as a verifier in math reasoning tasks to evaluate the correctness of a given answer (She et al., 2025; Gureja et al., 2025; Sun et al., 2025; Zhang et al., 2025). Based on the evaluation setup, reward models can be process-based, where the model assesses the reasoning step by step (called PRMs) (Lightman et al., 2024; Luo et al., 2024), while outcome reward models (ORMs) evaluate the entire reasoning (Cobbe et al., 2021; Shen et al., 2021; Hosseini et al., 2024; Zhang et al., 2024; Setlur et al., 2025). In this work, we focus on the latter and propose a novel cross-lingual outcome reward modeling framework that leverages complementary reasoning signals across languages.

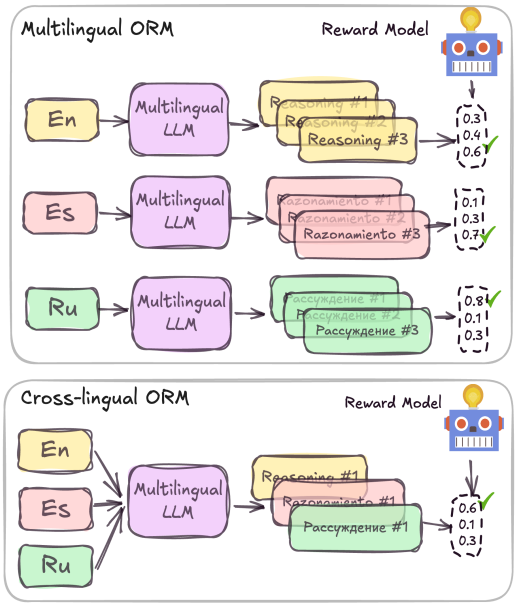


Figure 1: Illustration of multilingual ORM, where the verifier ranks responses within each language, and cross-lingual ORM (ours), where the verifier ranks responses across languages for the same question.

### 3.1 Cross-lingual Reward Modeling

Throughout this work, we distinguish between multilingual training and cross-lingual inference. In prior multilingual reward modeling approaches, a single verifier is trained on data from multiple languages, but candidate solutions are generated and ranked independently within each language. In contrast, our framework performs inference by generating candidate solutions in multiple languages for the same question and ranking them jointly across languages using a shared multilingual verifier.

To implement this idea, we extend the Best-of-N framework (Lightman et al., 2024) to a cross-lingual setting. Instead of sampling and ranking $N$ candidate solutions from a single language, we sample candidate solutions in multiple languages for the same question and select the highest-scoring solution across all languages using a trained verifier.

Given a math question $q$ and a generated candidate answer $a$, we train a discriminative verifier to predict whether the generated reasoning is correct. More specifically, we train an LLM as the verifier using the binary cross-entropy loss: $\mathcal{L}_{ORM} = -[y \cdot \log(\hat{y}) + (1-y) \cdot \log(1-\hat{y})]$. At inference, we use the verifier scores to rank a set of candidate answers in different languages for a given question using the probability that the model put on the correct class, and then we select

| | En. | Avg. | SC | Cross-SC | Multi-ORM | Cross-ORM | Pass@8-Multi | Pass@8-Cross |
|---|---|---|---|---|---|---|---|---|
| **Aya-Expanse-8b** | 79.6 | 63.4 | 58.3 | 79.7 | 73.3 | **83.2** | 82.4 | **93.2** |
| **Llama3.1-8b** | 80.4 | 64.0 | 71.6 | 73.4 | 76.2 | **84.0** | 86.9 | **92.4** |
| **Ministral-8b** | 82.0 | 65.1 | 70.3 | 78.1 | 76.4 | **87.6** | 84.3 | **93.4** |
| **Qwen2.5-7b** | 85.2 | 72.2 | 74.4 | 84.4 | 81.3 | **92.4** | 87.2 | **96.4** |
| **Phi3-7b** | 90.0 | 69.3 | 74.2 | 89.1 | 79.5 | **92.8** | 85.3 | **96.8** |
| **Llama3.2-3b** | 72.4 | 56.0 | 63.0 | 68.8 | 70.3 | **77.2** | 80.3 | **88.8** |

Table 1: The leftmost columns represent the English performance and the average performance of all the languages. *SC* denotes average self-consistency accuracy. *Cross-SC* represents the cross-lingual self-consistency baseline. *Pass@8-cross* outperforms the average *pass@8-multi*, indicating the complementary math reasoning skills across languages. Our proposed framework, *Crosslingual-ORM*, also exceeds the average *Multilingual-ORM* accuracy by a large margin.

the answer with the highest probability.

**Training Data Generation.** We use the Google Translate version of GSM8K training set in 8 languages, which includes around 7.5k examples per language of high-quality grade school math problems created by human writers (Cobbe et al., 2021) for our verifier training set (Lai and Nissim, 2024). We then prompt 3 models, the instruction-tuned version of Aya-Expanse 8B (Dang et al., 2024b), Llama3.1 8B (Grattafiori et al., 2024), and Qwen2.5 7B (Qwen Team, 2024), using the GSM8K training set in the covered languages to generate responses with step-by-step reasoning. We automatically labeled the generated reasoning paths as correct or incorrect based on the correctness of the final answer. Using generations from multiple models allows us to increase the size and diversity of the training set. To create a balanced dataset, we use the same number of correct and incorrect samples for each language, resulting in a set of around 88k samples for training.

**Cross-lingual ORM** We use the Qwen2.5-Instruct 3B model (Qwen Team, 2024) as the base of our reward model (verifier), as it has the widest officially supported language coverage among recent multilingual LLMs. We fine-tune it using the aforementioned training set[1]. To make a fair comparison, we use the same verifier for all experiments, including multilingual and cross-lingual ORM.

## 4 Experiments

### 4.1 Experimental setups

To study the chain-of-thought math reasoning ability of LLMs, we employ the MGSM (Multilingual Grade School Math) dataset (Shi et al., 2023). Following the original recipe of using MGSM (Shi et al., 2023), we prompt LLMs under the *Native-CoT* setting using 8-shots for all experiments, where the few-shot examples are in the same language as the question[2].

**Models.** We have carried out our analysis and experiments using a wide range of instruction-tuned multilingual models, including Aya-Expanse 8B (Dang et al., 2024b), Llama3.1 8B (Grattafiori et al., 2024), Qwen2.5 7B (Qwen Team, 2024), Ministral 8B[3], phi-3 7B (Abdin et al., 2024), and Llama 3.2 3b[4].

#### 4.1.1 Baselines.

We evaluate our cross-lingual ORM framework against the following baselines:

**Self-consistency.** A simple, yet effective approach in chain-of-thought (CoT) prompting is self-consistency (Wang et al., 2023; Yao et al., 2023). This baseline performs majority voting across a batch of sampled answers ($N = 8$) for each language, without any reward model.

**Cross-lingual Self-consistency.** We also consider a cross-lingual version of self-consistency as another baseline. In this setup, we do the majority voting across the generated answers across 8 languages (one answer per language). To have a robust evaluation, we repeat this experiment 8 times and report the average performance in Table 1.

**Multilingual-ORM.** Also known as *Best-of-N* technique, where the multilingual verifier scores $N$ different samples within a language and selects the

[1]Refer to the appendix A for details.

[2]Dataset details in the appendix A.

[3]https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

[4]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

one with the highest score (Wang et al., 2025). We use $N = 8$, generated with a temperature sampling of $T = 0.7$, and truncated at the top-p ($p = 0.95$) for all experiments (including the (cross-lingual) self-consistency baseline)[5]. We exclude the monolingual ORM baseline, where a separate RM is trained for each language and performs within-language sampling, as previous work reports that its performance is consistently lower than that of multilingual ORM (Wang et al., 2025).

### 4.2 Results and Findings

In the following, we analyze when and why cross-lingual ORM improves mathematical reasoning, focusing on the complementarity of reasoning knowledge across languages, the effect of sampling size, and the role of English on the cross-lingual ORM performance.

**LLMs exhibit complementary mathematical reasoning skills across languages.** To investigate the similarity of reasoning knowledge across languages, we employ pass@k, a well-established metric used to approximate the upper-bound performance of LLMs when generating multiple answers (Hosseini et al., 2024; Li et al., 2025b). This allows us to measure the degree of potential complementarity of reasoning knowledge between languages in multilingual LLMs. Pass@k considers a question solved if at least one answer in the sampled batch is correct, whether the batch spans multiple languages or a single language, depending on the experimental setup. In Table 1, we report pass@8 scores across languages (pass@8-Cross) and the average pass@8 scores across different samples within languages (pass@8-Multi).[6] We observe that Pass@8-Cross outperforms the performances of individual languages, suggesting that multilingual LLMs encode partially non-overlapping reasoning capabilities across languages, creating headroom for cross-lingual aggregation at inference time.

**Sampling across languages is superior to sampling within a language.** Building on our analysis, we employ the cross-lingual verifier described in Sec. 3.1 to see how languages benefit each other in practice. The middle part of Table 1 summarizes the accuracy of our cross-lingual ORM under within- and across-language settings. As shown, Cross-ORM clearly outperforms the average performance of ORM-Multi, with the largest benefits for non-English languages. These results suggest that leveraging cross-lingual signals is more effective than relying solely on monolingual reasoning, especially for underrepresented languages. The same pattern holds for both self-consistency baselines.

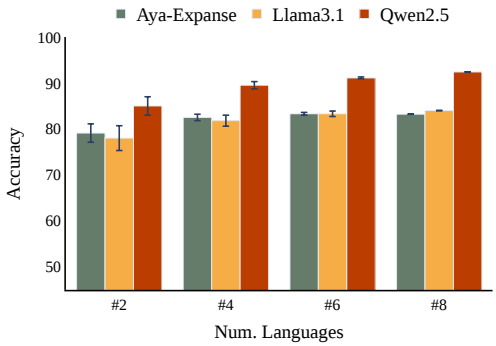


Figure 2: The mean and standard deviation cross-lingual ORM accuracy using different numbers of languages. Having more languages improves the performance.

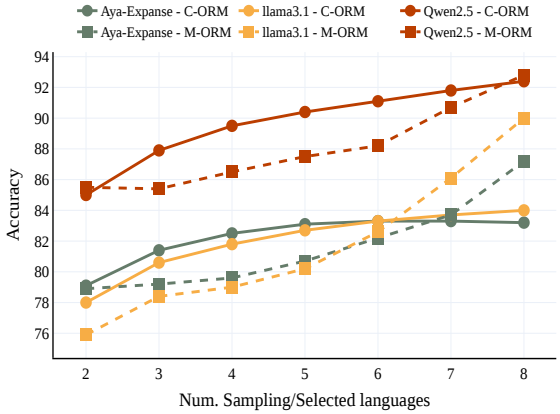


Figure 3: Comparing the accuracy of using cross-lingual ORM and multilingual ORM on English using the same number of languages and samples. Under smaller sampling budgets, cross-lingual ORM outperforms English performance.

**Increasing the pool of languages enhances the cross-lingual ORM performance.** To understand the impact of language pool size, we show the average performance for all possible language combinations at different pool sizes in Figure 2. As shown, the results demonstrate that adding more languages improves cross-lingual ORM performance up to a certain point, after which the additional gains become negligible.

[5]We abbreviate Cross-lingual and Multilingual as Cross and Multi in tables and figures for brevity.

[6]Per-language results are in the Appendix.

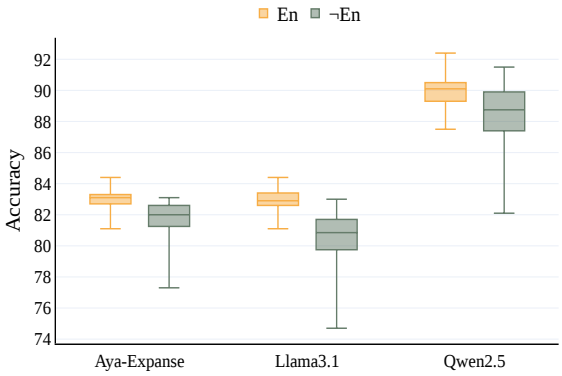


Figure 4: Average cross-lingual ORM performance across language pools of size 2–7, with and without English. English generally helps the reasoning performance, but some non-English sets outperform English-inclusive ones.

**Sampling across languages benefits English reasoning performance as well.** While our earlier analysis shows that cross-lingual ORM exceeds the average performance of multilingual ORM, its accuracy still lags behind that of English ORM. To better understand under what conditions other languages might benefit English, we compare the performance of English ORM (i.e., generating multiple answers in English) and cross-lingual ORM under different sampling budgets in Figure 3. Based on the results, we observe that cross-lingual ORM outperforms English ORM at low sampling budgets. However, this advantage fades as the number of samples increases. We suspect that additional sampling from other languages becomes redundant once English samples already cover a wide range of reasoning trajectories.

**Including English in language pools is generally helpful, yet it does not always lead to superior performance.** To examine the effect of including English in the language pools for the cross-lingual ORM setup, we report the mean and standard deviation of accuracy across all possible language pools with a size of 2 to 7 with and without English in Figure 4. As expected, including English generally improves cross-lingual ORM performance. However, this is not always the case; some language pools without English perform better than certain groups that include English, as reflected in the standard deviation of the non-English groups' performance.

## 5 Conclusion

In this paper, we present a novel cross-lingual reward modeling framework that effectively leverages complementary mathematical reasoning skills across languages in multilingual LLMs. Our experiments show that cross-lingual reward modeling outperforms its multilingual counterpart, benefiting even high-resource languages like English under low-budget inference settings. Furthermore, we find that languages mutually enhance each other's reasoning abilities. Together, these findings highlight inference-time cross-lingual aggregation as a simple yet effective mechanism for enhancing multilingual reasoning. Our results pave the way for future research into the similarities and differences of reasoning patterns across languages to improve multilingual reasoning models.

## 6 Limitations

A limitation of our work is that we focused solely on math reasoning tasks, and future research could explore other downstream tasks to broaden the applicability of our approach. Additionally, we used only eight languages, so expanding the number and diversity of languages would be important to further enhance our understanding of multilingual reasoning in LLMs. Another limitation is that we did not investigate the underlying reasoning patterns across languages, which could provide valuable insights for improving multilingual reasoning performance.

## 7 Acknowledgment

We want to thank the anonymous reviewers for their valuable comments and suggestions, which helped us in improving the paper. This research is funded in part by the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080 and in part by the European Union (ERC, CulturAL, 101171968). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

## References


Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

David Anugraha, Shou-Yi Hung, Zilu Tang, Annie En-Shiun Lee, Derry Tanti Wijaya, and Genta Indra Winata. 2025. mr3: Multilingual rubric-agnostic reward reasoning models. *Preprint*, arXiv:2510.01146.

Ju-Seung Byun, Jiyun Chun, Jihyung Kil, and Andrew Perrault. 2024. ARES: Alternating reinforcement learning and supervised fine-tuning for enhanced multi-modal chain-of-thought reasoning through diverse AI feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4410–4430, Miami, Florida, USA. Association for Computational Linguistics.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

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

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

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024b. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Srishti Gureja, Lester James Validad Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Triandi Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2025. M-RewardBench: Evaluating reward models in multilingual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 43–58, Vienna, Austria. Association for Computational Linguistics.

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, Albuquerque, New Mexico. Association for Computational Linguistics.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. 2024. V-STar: Training verifiers for self-taught reasoners. In *First Conference on Language Modeling*.

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

Jaedong Hwang, Kumar Tanmay, Seok-Jin Lee, Ayush Agrawal, Hamid Palangi, Kumar Ayush, Ila Fiete, and Paul Pu Liang. 2025. Learn globally, speak locally: Bridging the gaps in multilingual reasoning. *Preprint*, arXiv:2507.05418.

Miyoung Ko, Sue Hyun Park, Joonsuk Park, and Minjoon Seo. 2024. Hierarchical deconstruction of LLM reasoning: A graph-based framework for analyzing knowledge utilization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4995–5027, Miami, Florida, USA. Association for Computational Linguistics.

Huiyuan Lai and Malvina Nissim. 2024. mCoT: Multilingual instruction tuning for reasoning consistency

in language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12012–12026, Bangkok, Thailand. Association for Computational Linguistics.

Huiyuan Lai, Xiao Zhang, and Malvina Nissim. 2025. Multidimensional consistency improves reasoning in language models. *Preprint*, arXiv:2503.02670.

Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025a. Language models can easily learn to reason from demonstrations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15979–15997, Suzhou, China. Association for Computational Linguistics.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, Yingying Zhang, Fei Yin, Jiahua Dong, Zhijiang Guo, Le Song, and Cheng-Lin Liu. 2025b. From system 1 to system 2: A survey of reasoning large language models. *Preprint*, arXiv:2502.17419.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Jiayu Liu, Zhenya Huang, Chaokun Wang, Xunpeng Huang, ChengXiang Zhai, and Enhong Chen. 2025. What makes in-context learning effective for mathematical reasoning. In *Forty-second International Conference on Machine Learning*.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, and Abhinav Rastogi. 2024. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Leonardo Ranaldi and Andre Freitas. 2024. Self-refine instruction-tuning for aligning reasoning in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2325–2347, Miami, Florida, USA. Association for Computational Linguistics.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2025. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *The Thirteenth International Conference on Learning Representations*.

Shuaijie She, Junxiao Liu, Yifeng Liu, Jiajun Chen, Xin Huang, and Shujian Huang. 2025. R-PRM: Reasoning-driven process reward modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13438–13451, Suzhou, China. Association for Computational Linguistics.

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

Jianhao Shen, Yichun Yin, Lin Li, Lifeng Shang, Xin Jiang, Ming Zhang, and Qun Liu. 2021. Generate & rank: A multi-task framework for math word problems. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2269–2279, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Wei Sun, Qianlong Du, Fuwei Cui, and Jiajun Zhang. 2025. An efficient and precise training data construction framework for process-supervised reward model in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4292–4305, Vienna, Austria. Association for Computational Linguistics.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025. Demystifying multilingual chain-of-thought in process reward modeling. *arXiv preprint arXiv:2502.12663*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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

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

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

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024. Generative verifiers: Reward modeling as next-token prediction. In *The 4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10495–10516, Vienna, Austria. Association for Computational Linguistics.

# A Experimental Setups

## A.1 Cross-lingual ORM setups

We fine-tune our verifier for 5 epochs, with AdamW, a learning rate of $2e$-4, and a batch size of 96. Since the main task is binary classification and to make fine-tuning efficient, we fine-tune with LoRA (Hu et al., 2022) with a rank of 16 and scaling factor of 32.

The MGSM dataset covers 11 languages, including English(en), Spanish(es), French(fr), German(de), Russian(ru), Chinese(zh), Japanese(ja), and Thai(th), with 250 examples for each. We exclude Swahili, Telugu, and Bengali, as multilingual LLMs perform poorly on these languages. Thus, training a reward model (verifier) that already does not have acceptable linguistic knowledge on certain languages for a reasoning task is not practical. Nevertheless, our selection includes languages from diverse language families and writing scripts (Lai et al., 2025). We have used the evaluation harness framework for our experiments and reported exact-match scores (Gao et al., 2024) .

# B Full Results

**Question:**

Melanie is a door-to-door saleswoman. She sold a third of her vacuum cleaners at the green house, 2 more to the red house, and half of what was left at the orange house. If Melanie has 5 vacuum cleaners left, how many did she start with?

**English Response:**

Melanie sold a third of her vacuum cleaners at the green house. That means she sold 1/3 of her vacuum cleaners. 1/3 of 5 is 5/3, which is 1.67. Melanie sold 1.67 vacuum cleaners at the green house. She had 5 left, so she had 5 + 1.67 = 6.67 vacuum cleaners before the green house. She sold 2 more at the red house, so she had 6.67 - 2 = 4.67 vacuum cleaners before the red house. She sold half of what was left at the orange house. Half of 4.67 is 2.335. Melanie had 4.67 - 2.335 = 2.335 vacuum cleaners left. Melanie had 5 vacuum cleaners left, so she must have had 5 + 2.335 = 7.335 vacuum cleaners before the orange house. Melanie had 7.335 vacuum cleaners before the orange house. The answer is 7.335.

**Japanese Response:**

緑の家で3分の1の掃除機を売った場合、3/1*3=9個の掃除機を売ったことになります。赤の家で2台を売ったので、合計9+2=11台を売ったことになります。オレンジの家で残りの半分を売った場合、残りの半分は 1/2*5=2.5 となり、オレンジの家で 2.5 台を売ったことになります。合計で、メラニーは 11+2.5=13.5 台の掃除機を売りました。掃除機は整数でなければならないため、彼女は 13 台の掃除機を売りました。彼女は 5 台の掃除機を手元に残したため、彼女は 13+5=18 台の掃除機を持っていました。答えは18です。

**Google translation of Japanese Response:**

If she sells one-third of the vacuums at the green house, she has sold 3/1*3=9 vacuums. She sells two at the red house, so she has sold a total of 9+2=11 vacuums. If she sells the other half at the orange house, the remaining half is 1/2*5=2.5, so she has sold 2.5 at the orange house. In total, Melanie has sold 11+2.5=13.5 vacuums. Because vacuums must be whole numbers, she has sold 13 vacuums. She has 5 vacuums left, so she has 13+5=18 vacuums. The answer is 18.

Figure 5: A cherry-picked example of Llama 3.1's responses to the same question in English and Japanese, where the English answer is incorrect but the Japanese response is correct, illustrating the complementary reasoning skills across languages.

| | en | fr | es | de | ru | zh | ja | th | avg. |
|---|---|---|---|---|---|---|---|---|---|
| **Aya-Expanse-8b** | | | | | | | | | |
| CoT | 76.6 ±1.1 | 65.1 ±2.0 | 73.0 ±1.8 | 68.8 ±1.9 | 67.6 ±2.5 | 63.7 ±0.8 | 57.7 ±1.2 | 19.7 ±1.8 | 61.5 |
| SC | 84.0 | 70.8 | 76.8 | 76.8 | 73.6 | 71.2 | 67.2 | 23.2 | 58.3 |
| **Llama3.1-8b** | | | | | | | | | |
| CoT | 75.0±1.7 | 60.4±1.4 | 66.8±1.9 | 59.6±3.8 | 61.1±2.8 | 57.5±2.3 | 46.9±2.6 | 47.4±1.1 | 59.4 |
| SC | 84.8 | 72.4 | 80.0 | 74.0 | 74.0 | 69.6 | 58.8 | 59.2 | 71.6 |
| **Ministral-8b** | | | | | | | | | |
| CoT | 77.4±1.5 | 65.1±0.7 | 71.4±1.3 | 64.9±1.8 | 65.4±0.8 | 58.8±1.7 | 43.8±1.9 | 44.5±1.1 | 61.4 |
| SC | 87.2 | 72.0 | 80.0 | 74.0 | 75.2 | 67.6 | 55.2 | 54.0 | 70.3 |
| **Qwen2.5-7b** | | | | | | | | | |
| CoT | 84.3±1.0 | 70.2±1.8 | 76.3±2.3 | 63.8±1.2 | 69.0±1.7 | 67.8±1.7 | 63.7±1.4 | 51.8±1.3 | 68.4 |
| SC | 89.2 | 73.6 | 82.8 | 70.8 | 73.6 | 75.6 | 70.0 | 59.6 | 74.4 |
| **Phi3-7b** | | | | | | | | | |
| CoT | 87.6±1.8 | 77.0±2.3 | 83.7±1.7 | 77.3±0.9 | 71.6±1.8 | 69.2±2.3 | 58.0±1.4 | 18.2±1.3 | 67.8 |
| SC | 92.8 | 84.4 | 88.4 | 83.6 | 82.8 | 74.4 | 65.6 | 23.6 | 74.2 |
| **Llama3.2-3b** | | | | | | | | | |
| CoT | 65.6±1.1 | 51.0±2.7 | 56.4±1.0 | 52.4±1.1 | 53.1±2.4 | 48.8±2.1 | 32.0±1.7 | 44.1±1.9 | 50.4 |
| SC | 79.2 | 60.4 | 70.8 | 67.2 | 65.6 | 62.8 | 44.4 | 55.6 | 63.0 |

Table 2: Vanilla Chain-of-thought(CoT) performance and self-consistency (SC) on MGSM.

| | en | fr | es | de | ru | zh | ja | th | avg. |
|---|---|---|---|---|---|---|---|---|---|
| Aya-Expanse-8b | 94.4 | 85.6 | 90.8 | 89.2 | 87.6 | 84.4 | 82.4 | 44.4 | 82.3 |
| Llama3.1-8b | 94.8 | 85.2 | 92.4 | 89.2 | 92.0 | 86.8 | 74.8 | 80.0 | 86.9 |
| Ministral-8b | 94.4 | 85.6 | 91.6 | 85.2 | 86.4 | 86.4 | 73.2 | 71.6 | 84.3 |
| Qwen2.5-7b | 95.6 | 84.0 | 94.4 | 85.2 | 87.6 | 91.2 | 84.0 | 75.6 | 81.3 |
| phi3-3b | 96.4 | 92.0 | 95.6 | 90.8 | 92.0 | 88.4 | 84.8 | 42.0 | 85.3 |
| Llama3.2-3b | 91.6 | 78.8 | 85.6 | 83.2 | 82.8 | 80.8 | 66.0 | 73.6 | 80.3 |

Table 3: Comparison of Pass@8-Multi across different languages on the MGSM task.

| | en | fr | es | de | ru | zh | ja | th | avg. |
|---|---|---|---|---|---|---|---|---|---|
| Aya-Expanse-8b | 87.2 | 76.4 | 83.2 | 76.8 | 78.4 | 76.0 | 71.2 | 37.2 | 73.3 |
| Llama3.1-8b | 90.0 | 77.6 | 84.0 | 60.4 | 80.0 | 78.8 | 68.4 | 70.0 | 76.2 |
| Ministral-8b | 91.2 | 76.0 | 87.6 | 77.6 | 77.2 | 78.0 | 60.4 | 62.8 | 76.4 |
| Qwen2.5-7b | 92.8 | 80.0 | 86.4 | 74.8 | 82.4 | 85.2 | 76.0 | 72.4 | 81.3 |
| Phi3-7b | 94.0 | 86.8 | 91.2 | 86.0 | 84.0 | 83.2 | 75.6 | 35.2 | 79.5 |
| Llama3.2-3b | 83.2 | 70.0 | 76.4 | 62.4 | 75.2 | 70.8 | 56.8 | 67.2 | 70.3 |

Table 4: Comparison of Multi-ORM across different languages on the MGSM task.

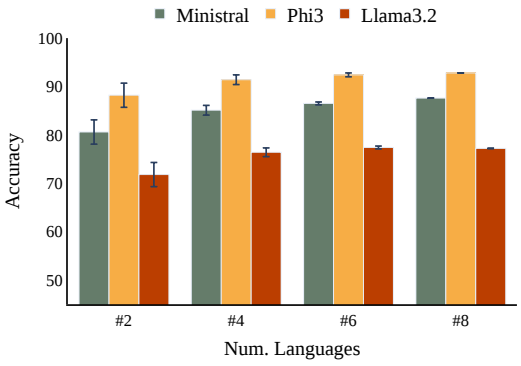


Figure 6: The average and standard deviation cross-lingual ORM performance using different numbers of languages.

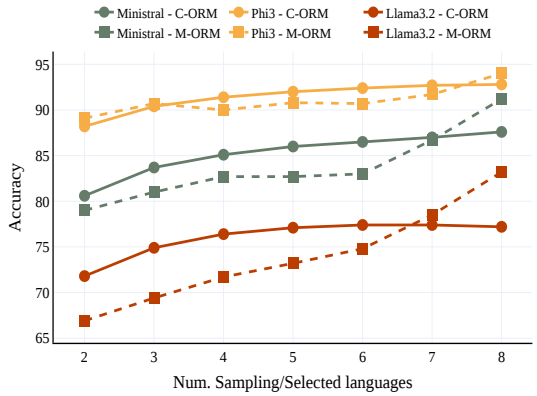


Figure 7: Comparing the accuracy of cross-lingual ORM and multilingual ORM of English using the same number of languages and samples.

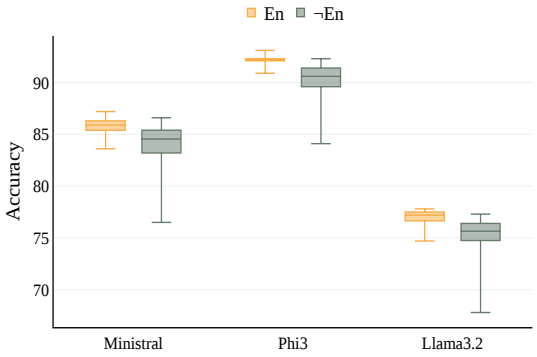


Figure 8: Comparing the role of English on the performance.