Unlocking Multilingual Math Potential: Strategies Amidst Data Scarcity

Anonymous ACL submission

Abstract

Existing works studying behaviors of Large Language Models (LLMs) in multilingual settings focus mainly on general downstream tasks such as instruction following. To fill this gap, we perform an in-depth analysis of LLMs' math reasoning capacity under multilingual settings and propose to alleviate the shortage of high quality multilingual math reasoning post-training data by exploring whether prior English math knowledge and additional English data helps, and by observing the effects of multilingual synthetic data on performance. For models pre-trained with mostly English data, we find that prior English math knowledge helps and that scaling English data helps only when the training and evaluation data belong to similar distributions (human/machine translated). Additonally, we find that inclusion of multilingual synthetic data leads to improved performance on human-translated data but degraded performance on machine-translated data. Our findings shed light on effective finetuning of LLMs for better multilingual math reasoning performance given the shortage high-quality multilingual math reasoning data.

1 Introduction

In the era of modern Large Language Models (LLMs), the rising interest in multilingual capabilities of models has led to the extension of existing datasets into different languages via human translation (Shi et al., 2023), machine translation services or LLM-based translation (Chen et al., 2023b; Li et al., 2023; Lai et al., 2023; Chai et al., 2024; Lai and Nissim, 2024). Detailed analyses for behaviors of models on translated datasets exists for tasks like multilingual instruction following (Shaham et al., 2024), but are missing for multilingual math reasoning, which is an important task to measure the reasoning capabilities of LLMs. Moreover, the low availability of multilingual math training data, as compared to English, hinders improvements as well. In this work, we study model behaviors and explore various strategies to improve multilingual math reasoning performance given the scarcity of high quality multilingual data.

We first study the variance of cross-lingual performance for a wide pool of LLMs when post-trained for a certain language on math reasoning task. We observe that finetuning multilingual LLMs (which are already pre-trained with target task multilingual data) with additional math reasoning data belonging to one language does not lead to highly varying performance difference between the training language and other languages, indicating that post-training in this scenario leads to saturating results. This is not the case for monolingual models (pretrained mostly with English data), which exhibit high performance variance across languages.

Despite their variance, monolingual models are suitable to study the impact of different training and data setups for multilingual tasks. We observe that monolingual models having math reasoning capabilities in English transfer better to multilingual setting compared to the base model, implying that intermediate training with additional English math reasoning data helps in overall performance. Further, we find that scaling English training data while keeping non-English data constant benefits performance on evaluation sets which align with the training data in-terms of dataset construction (human v/s machine-translated).

The above disparity between data distribution of evaluations datasets and the gaining popularity of synthetic data usage for post-training (Huang et al., 2023; Sun et al., 2024; Ri et al., 2024; Dubey et al., 2024), prompt us to examine the effect of introduction of multilingual synthetic data on available benchmarks. We find that scaling synthetic multilingual data obtained via self-training helps the model to consistently perform better on human-translated test set while leading to diminishing performance on machine-translated test set. This result highlights the importance of determining the optimal mix of multilingual datasets (which are dominantly machine-translated) with synthetic data for obtaining maximal model performance.

Overall, our experiments provide an in-depth understanding of the status of multilingual math reasoning, along with insights that can be leveraged to train better models. More specifically, models with prior English math knowledge and determining the appropriate proportion of English and multilingual synthetic data in the training data play are crucial for multilingual math reasoning.

2 Related Work

Multilingual Math Reasoning Measuring the capability of models to solve mathematical problems is directly correlated with their complex reasoning proficiency. Due to the lack of abundant multilingual domain data, most works rely on extending datasets using translation models or frontier LLMs (Chen et al., 2023a; Shi et al., 2023; Lai and Nissim, 2024). Works have also focused on leveraging auxilary tasks to better align models for multilingual data (Zhu et al., 2024a,b). Another line of works emphasize on studying the reasoning consistency of LLMs in different languages Shi et al. (2023); Chowdhery et al. (2023); Lai and Nissim (2024). However, none of the works study the differences between the behaviors of monolingual and multilingual models when finetuned for multilingual math.

Influence of data attributes on multilingual performance Cross-lingual transfer is a well known phenomena, which has been shown to be effective in cases with limited multilingual data (Shaham et al., 2024; Chen et al., 2023b). Prior works have shown that multilingual performance benefits from intermediate multitask training in English (Phang et al., 2020) and from an imbalance proportion of languages in training data (Schäfer et al., 2024). While these works are aimed at studying the impact of data for multilingual tasks in general, our work is focused on exploring the effects specifically in math settings.

The predominant focus of related works has been disjointly on multilingual math reasoning and the impact of data attributes for multilingual tasks in general. In this paper, we address the gap of an exhaustive comparison of monolingual and multilingual models for multilingual math reasoning and explore techniques to maximize performance by leveraging English and multilingual synthetic data.

3 Task Setup

The task is framed as a supervised finetuning task, where given a math problem, the objective is to generate the steps to solve the problem and output the final answer. In addition to the question, the model is also prompted with the Chain of Thought prompt template (in the language corresponding to the question), which has been corroborated to enhance performance (Wei et al., 2022). We use the mCoT dataset (Lai and Nissim, 2024) as the training data for all our experiments and use answer exact match (accuracy) as our evaluation metric on the MGSM (Shi et al., 2023) and MSVAMP (Chen et al., 2023a) datasets.

4 Improvement strategies under data constraint settings

4.1 Cross-lingual performance variance for math/reasoning when trained on a single language

We first explore the cross-lingual performance variance in LLMs when trained on a single language. We perform full finetuning of the models for only one epoch to avoid any possibility of overfitting.

As shown in Figure 2, we observe that multilingual models like Aya23 (Aryabumi et al., 2024) and Qwen2 (Yang et al., 2024), when trained on different languages exhibit similar performance (low variance) for any particular language, though the performance range varies depending on the model and the resource level of the language. This implies that the models rely on their existing knowledge despite being post-trained specifically for a particular language. We hypothesize that the additional training with data belonging to a particular language does not influence the model much since it has already been trained on vast amounts of multilingual data during the pre- and post- training stages.

These observations are contrary for models that have a low proportion of multilingual data in their pre-training corpus like Mistral-7B (Jiang et al., 2023a) and Llama3-8B (Dubey et al., 2024) (Figure 3). Such models exhibit considerable variance when the evaluation data language is constant due to additional learning signals provided to them.

Next, we study the correlations between languages. Specifically, we tabulate the languages (in Figure 4) which when used to train models lead to the top-k performance in a particular language. We consider only Mistral-7B-v0.1 Base and Llama3 Base for computing the results since their monolingual nature provides the best setup to study the influence of languages on performance. We consider evaluation metrics on the MGSM and MSVAMP datasets to bolster the confidence of the observations and set $k = 4$. Unsurprisingly, we observe that the target language benefits from training data in the same language and that European languages (English, Spanish, French, and German) help each other.

Based on the above results, monolingual models serve the optimal setting to study the influence of different training strategies and dataset mixtures since multilingual models might not be affected by them.

4.2 Does additional English math/reasoning data help in multilingual setting?

LLMs post-trained for a specific task or data might locate the model in a local minima, which leads to poor adaptation and generalization for new tasks/data. We intend to explore if this is the case when LLMs with math reasoning capabilities in English are adapted to multilingual data. To make fair comparisons, we experiment with WizardMath (Luo et al., 2023), which is a Mistral-7B-v0.1 Base model, trained for English Math problems using Reinforcement Learning, and Mathstral [1], another variant of the Mistral base model. We follow training hyperparameters similar to Lai and Nissim (2024) and train the models for two epochs.

We observe that both WizardMath and Mathstral outperform Mistral-7B-v0.1 (Table 1) with improvements on almost all languages in the evaluation data (Figure 5). We hypothesize two reasons for this gain in performance namely, additional English training data [2] and training strategies other than supervised finetuning (Uesato et al., 2022; Luo et al., 2023; Lightman et al., 2023; Gao et al., 2024)

Continuing our previous discussion and given that English data is easily available for most tasks, we design experiments in a controlled setting to verify if additional English training data leads to improved performance. We sample 25% of non-English data from our training set and gradually increase the proportion of English data. Thus, we sample 25%, 50%, 75% and 100% of English data and mix it with the sampled non-English data.

[1] https://mistral.ai/news/mathstral/

[2] The models have benchmarks on GSM 8K only, which suggests the model being suitable mainly for English

Finetuned Model	MGSM	MSVAMP
Mistral-7B-v0.1	0.68	0.668
WizardMath	0.707	0.696
Mathstral	0.748	0.752

Table 1: Average accuracy of English Math Reasoning models in multilingual setting. **Takeaway**: English math knowledge improves performance

Figure 1 shows the results of training a Mistral-7B-v0.1 model on these 4 data mixtures on the MGSM and MSVAMP datasets. We observe that scaling English data mostly improves performance in English with slight effects on non-English data, though the overall performance variations are negligible on MGSM, while non-trivial on MSVAMP. We hypothesize that this is due to data belonging to different distributions. Scaling English data whose Google translations constitute the multilingual component of the training set helps in improving performance on MSVAMP (Google Translated) and does not affect results on MGSM (human translated).

Thus, training models with prior math knowledge is beneficial for multilingual math reasoning, though sole reliance on machine-translated data must be avoided as it might only help improvements on machine-translated benchmarks and lead to plateauing performance on human-translated ones.

4.3 Impact of synthetic multilingual data

We create another controlled setting at a small scale to explore the influence of synthetic multilingual data. We sample 5K samples belonging to each of the 11 languages from the mCoT training data. We call this dataset *mCoT-55K*. We follow the procedure devised by Wang et al. (2024) to incorporate synthetic data into our training pipeline but restrict it to one iteration of synthetic data generation. First, we perform supervised finetuning (SFT) of Mistral-7B-v0.1 on *mCoT-55K*, which is treated as the baseline for the experiment. Then we sample 5 generations for each sample from the SFT model at a temperature of 0.7. Using the ground truth answer, we construct pair-wise data to train the SFT model further using Direct Policy Optimization (Rafailov et al., 2024) (DPO). The DPO model is then used to sample positive samples, which are then included with the *mCoT-55K* samples (we use the same sampling settings as earlier). The SFT model is then trained on this combination of origi-

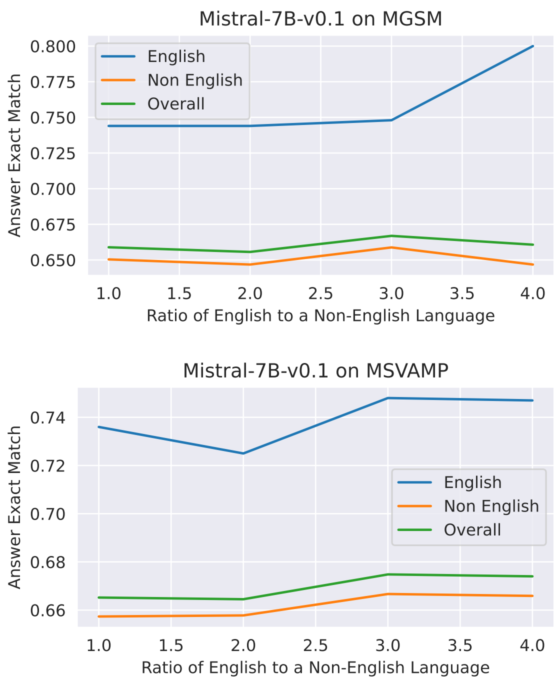

Figure 1: Mistral-7B-v0.1 on constant non-English samples and varying proportion of English samples. **Takeaway**: Scaling proportion of English helps when training and test data are created in a similar manner

nal training data and synthetic data to get the final model.

We ablate on the proportion of synthetic data used in the final training. More specifically, we gradually add 25%, 50%, 75% and 100% of the generated synthetic SFT data ensuring a balance in the samples per language (entire synthetic data has 3100 samples per language).

We find that the addition of synthetic data always achieves better performance than the baseline model on the human-translated MGSM test set, whereas the performance on the Google-translated MSVAMP is always poor compared to the baseline (Table 2). For MGSM, we do not find any particular trend as the synthetic data is scaled, which can be attributed to different synthetic samples having different quality levels. For MSVAMP, we find that the performance decreases as the proportion of synthetic data increases. We hypothesize that the addition of synthetic data adds diversity to the Google-Translated training dataset and shift the overall data distribution away from Google-Translated distribution, thus leading to degrading performance on MSVAMP and improved results on MGSM.

The above observations signify that the an appropriate mixture of machine-translated data and multilingual synthetic data is mandatory to optimize performance on both human-translated and machine-translated evaluation benchmarks and additional work is needed to devise methods and provide recommendations for the same.

Dataset	MGSM	MSVAMP
mCoT-55K	0.469	0.615
+ 25% synthetic	0.532	0.596
+ 50% synthetic	0.515	0.577
+ 75% synthetic	0.517	0.553
+ 100% synthetic	0.497	0.525

Table 2: Results for multilingual synthetic data scaling. **Takeaway**: Synthetic data is similar to human-generated data since it leads to consistent improvements on human-translated test bed (MGSM), but degradation on Google-Translated test bed (MSVAMP)

5 Conclusion

We study various aspects revolving around multilingual math settings regarding the performance variance of monolingual and multilingual LLMs when trained for only one language, the transfer of English math reasoning models to multilingual data, and the impact of multilingual synthetic data. Given that the majority of multilingual task data is machine-translated, determining the optimal proportion of multilingual synthetic data becomes crucial to align models toward human generated data. Overall, our findings help us understand the behaviors and training dynamics of various models in different training data settings and thus help in designing better experimental setups for improving multilingual math reasoning performance.

6 Limitations

First, our study is limited to 11 languages, for which original English datasets were translated. Moreover, amongst the 11 languages, very few of them belong to the low-resource category. Second, the training dataset used is constructed entirely using machine-translation (Google Translate). Prior works have highlighted the difference between machine-translated data and human-translated data (Jiang et al., 2023b; Luo et al., 2024). Further, we experiment with only one technique of synthetic data generation and restrict our study to models with parameter count in the 7-8 B range due to computation constraints.

References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, et al. 2024. xcot: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning. *arXiv preprint arXiv:2401.07037*.

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

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Barry Haddow, and Kenneth Heafield. 2023b. Monolingual or multilingual instruction tuning: Which makes a better alpaca. *arXiv preprint arXiv:2309.08958*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

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

Bofei Gao, Zefan Cai, Runxin Xu, Peiyi Wang, Ce Zheng, Runji Lin, Keming Lu, Junyang Lin, Chang Zhou, Tianyu Liu, et al. 2024. The reason behind good or bad: Towards a better mathematical verifier with natural language feedback. *arXiv preprint arXiv:2406.14024*.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhaokun Jiang, Qianxi Lv, and Ziyin Zhang. 2023b. Distinguishing translations by human, nmt, and chatgpt: A linguistic and statistical approach. *arXiv preprint arXiv:2312.10750*.

Huiyuan Lai and Malvina Nissim. 2024. mcot: Multilingual instruction tuning for reasoning consistency in language models. *arXiv preprint arXiv:2406.02301*.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

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

Jiaming Luo, Colin Cherry, and George Foster. 2024. To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation. *Transactions of the Association for Computational Linguistics*, 12:355–371.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575, Suzhou, China. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Ryokan Ri, Shun Kiyono, and Sho Takase. 2024. Self-translate-train: A simple but strong baseline for cross-lingual transfer of large language models. *arXiv preprint arXiv:2407.00454*.

Anton Schäfer, Shauli Ravfogel, Thomas Hofmann, Tiago Pimentel, and Imanol Schlag. 2024. Language imbalance can boost cross-lingual generalisation. *arXiv preprint arXiv:2404.07982*.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. *arXiv preprint arXiv:2401.01854*.

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*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2024. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Tianduo Wang, Shichen Li, and Wei Lu. 2024. Self-training with direct preference optimization improves chain-of-thought reasoning. *arXiv preprint arXiv:2407.18248*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Andrea W Wen-Yi, Unso Eun Seo Jo, Lu Jia Lin, and David Mimno. 2024. How chinese are chinese language models? the puzzling lack of language policy in china's llms. *arXiv preprint arXiv:2407.09652*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jun Zhao, Zhihao Zhang, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *arXiv preprint arXiv:2401.01055*.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. 2023. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*.

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

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024b. Question translation training for better multilingual reasoning. *arXiv preprint arXiv:2401.07817*.

A Experimental Settings

We train our models using a learning rate of $5e-6$ and a batch size of 8 per GPU with 16 steps of gradient accumulation. Cosine learning rate scheduler is used with a linear warm up of 3% training steps. The model is trained for a maximum sequence length of 1024 and the number of training epochs is set to 1 and 2 for the language-wise and all language training. We use Pytorch (Paszke et al., 2019), Huggingface's transformers (Wolf et al., 2020) and native Pytorch Fully Sharded Data Parallel (FSDP) (Zhao et al., 2023) on 8 x H100s for training our models.

B Comparison of models when trained on a single language

Figures 2, 3 and 4 show the results for Section 4.1. An interesting observation for Llama3-8B Base is that training on Chinese (high resource) and Thai (low resource) does not lead to the highest performance. Many works revolve around adapting Llama-based models on Chinese data (Zhao et al., 2024; Cui et al., 2023). On the other hand, Wen-Yi et al. (2024) show that the behavior of international LLMs is similar to Chinese LLMs. Thus, future work is needed to investigate if this trend exhibited by Llama3 holds for other domains and tasks.

C Language Wise Results for Mistral variants

Figure 5 shows the results for Section 4.2

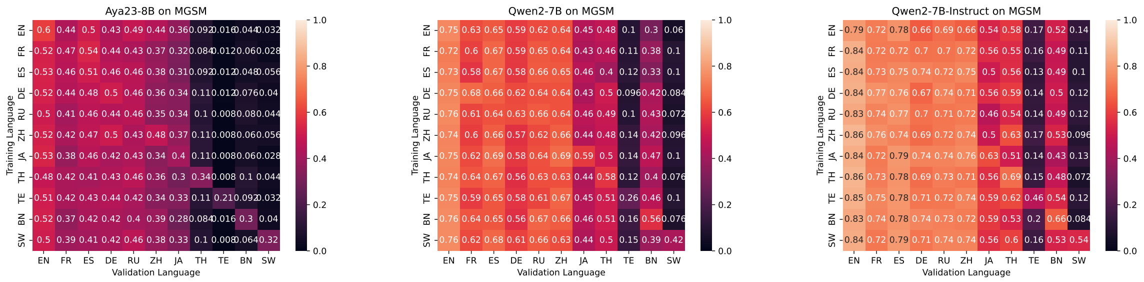

Figure 2: Language-wise performance matrix for multilingual models on MGSM. **Takeaway**: Performance on a fixed evaluation language is consistent for multilingual models despite a language specific post-training

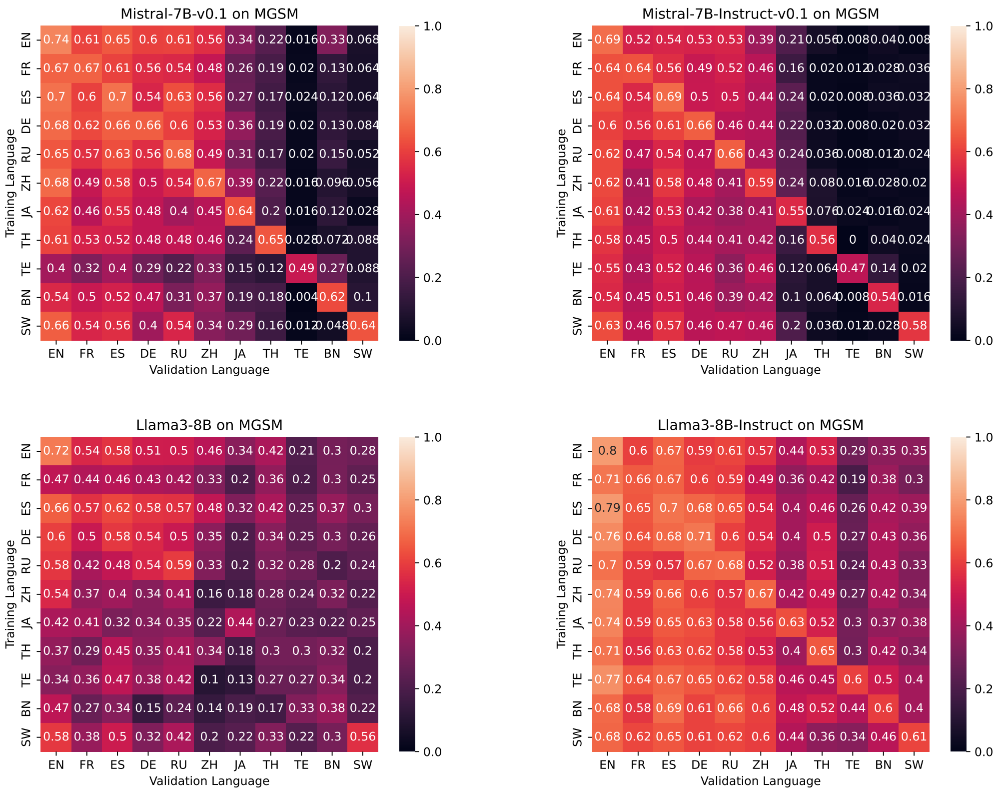

Figure 3: Language-wise performance matrix for monolingual models on MGSM. **Takeaway**: Performance on a fixed evaluation language varies significantly as compared to multilingual models

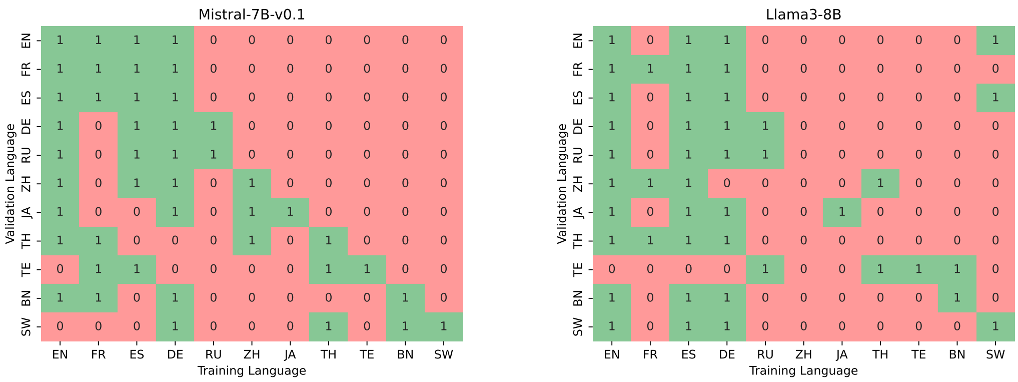

Figure 4: Training Languages which lead to highest performances while validating for a particular language. **Takeaway**: Same training and validation language work better in most cases; some exceptions are found for Llama3-8B Base

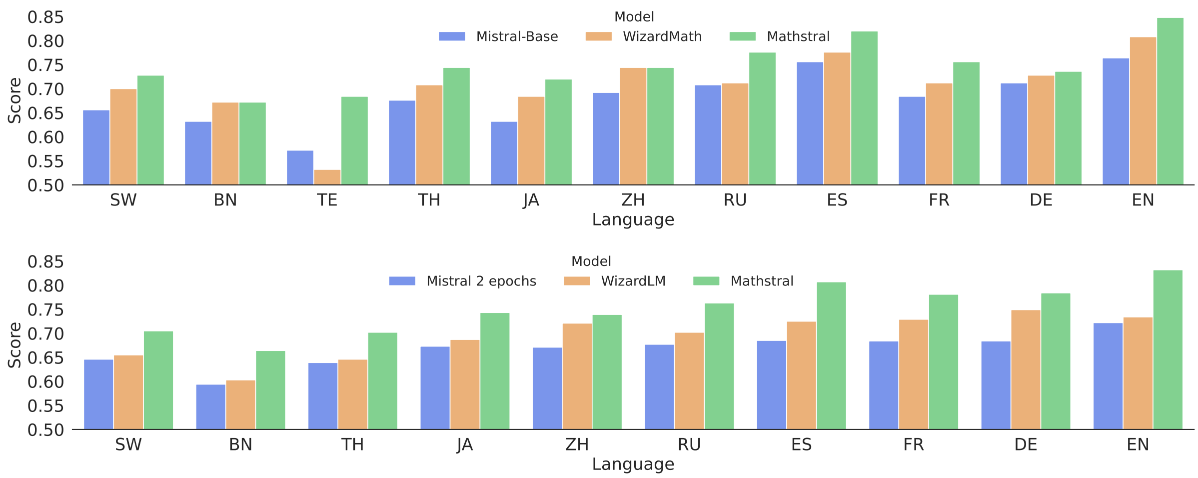

Figure 5: Evaluation of Mistral variants on MGSM (top) and MSVAMP (bottom) **Takeaway**: English math knowledge improves performance