

the correct answer to the query in high-resource languages like English, they struggle with the same query in low-resource languages. These discrepancies emphasize a fundamental limitation in the understanding and reasoning capabilities of LLMs for low-resource languages.

To alleviate this issue, we propose a two-stage instruction tuning framework (LinguaLIFT) for enhancing low-resource language tasks. An additional language alignment layer is initially integrated into the LLM to adapt a pre-trained multilingual encoder. LinguaLIFT then improves multilingual alignment via code-switched fine-tuning, using the code-switched translation data generated from unsupervised multilingual alignment lexicons. The second stage fine-tunes LLMs with English-only instruction data while freezing the language alignment layer, which enables LLMs to transfer the learned task-specific capabilities from English to low-resource languages. Additionally, we introduce a new benchmark named Multilingual Math World Problem (MMWP), spanning 21 low-resource, 17 medium-resource, and 10 high-resource languages to evaluate multilingual reasoning tasks comprehensively. Experiments demonstrate that LinguaLIFT significantly outperforms several strong competitive methods on the MMWP and other widely used benchmarks, such as MGSM (Shi et al., 2023), MSVAMP (Chen et al., 2023), XNLI (Conneau et al., 2018) and X-CSQA (Lin et al., 2021).

2 Related Work

Leveraging Powerful Cross-lingual Transfer Capabilities of Multilingual Models. Recent studies have shown that multilingual models possess strong cross-lingual transfer capabilities (Conneau et al., 2020; Xue et al., 2021; FitzGerald et al., 2023; Chirkova and Nikoulina, 2024; Shaham et al., 2024; Chen et al., 2024c), benefiting the low-resource language community by enabling task transfer across multiple languages after fine-tuning on high-resource languages (Kew et al., 2023; Pfeiffer et al., 2020; Reimers and Gurevych, 2020; Pan et al., 2021; Feng et al., 2022).

Unlike previous approaches focusing on multilingual instruction tuning (Singh et al., 2024; Üstün et al., 2024; Köpf et al., 2024; Zhu et al., 2024b,a; Li et al., 2023a; Wei et al., 2023; Ranaldi et al., 2023), LinguaLIFT incorporates a language alignment layer into LLMs to adapt a pre-trained

multilingual encoder through code-switched fine-tuning. While freezing the language alignment layer and a pre-trained multilingual encoder, the LLM is fine-tuned with English-only instruction data to transfer task-specific capabilities from English to low-resource language tasks.

Improving Multilingual Mathematical Reasoning Tasks. Recent efforts to improve multilingual mathematical reasoning for LLMs can be categorized into three ways: 1) **Prompting close-source LLMs:** Qin et al. (2023), Shi et al. (2023), and Huang et al. (2023) designed prompts for closed-source LLMs like ChatGPT, translating non-English contexts into English for reasoning. However, this approach is limited by translation quality and does not improve multilingual understanding or work well for open-source LLMs (Zhu et al., 2024b). 2) **Instruction-tuning open-source LLMs:** Chen et al. (2023), Chai et al. (2024), and Lai and Nissim (2024) adopted a translate-training method, translating English reasoning datasets into non-English and instruction-tuning LLMs. Zhu et al. (2024b) and She et al. (2024) proposed approaches to transfer mathematical reasoning capabilities from English to non-English. While these methods improve multilingual reasoning, they incur high translation costs and errors, making them impractical for low-resource languages. 3) **Bridging existing skilled LLMs to multilingualism:** Yoon et al. (2024) combines pre-trained multilingual models with skilled reasoning LLMs, but a performance gap persists between low-resource and high-resource languages. While Huang et al. (2024) shows significant improvement, it relies heavily on parallel and multilingual task data.

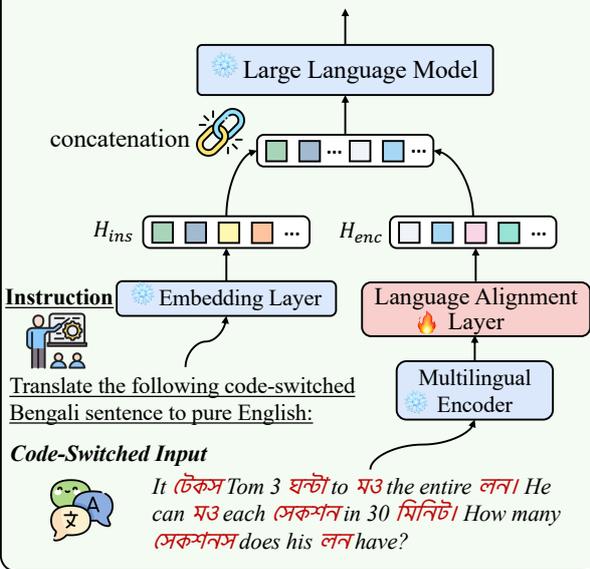
Different from existing studies, this paper proposes a novel two-stage instruction fine-tuning framework to enhance the reasoning capabilities of low-resource language tasks without relying on multilingual instruction data.

3 Methodology

Figure 2 illustrates an overview of the proposed two-stage instruction tuning framework, LinguaLIFT. The core idea is to transfer task-specific capabilities learned from English to low-resource languages, leveraging the language alignment established by code-switched tuning. We then introduce the model architecture (§3.1), the language alignment stage (§3.2) and the task transfer stage (§3.3).

Stage-I: Language Align

It takes Tom 3 hours to mow the entire lawn. He can mow each section in 30 minutes. How many sections does his lawn have?



Stage-II Task Transfer

To find out the number of sections Tom's lawn has, we need to determine how many 30-minute intervals are in 3 hours. There are 60 minutes in 1 hour. So, 3 hours will have $3 * 60 = 180$...

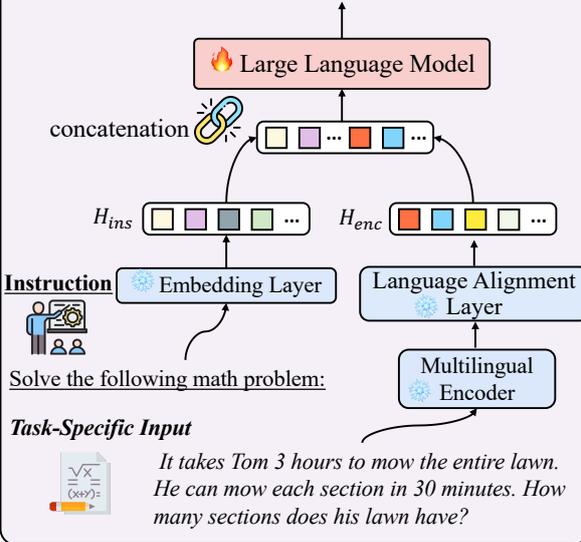


Figure 2: Overview of the proposed LinguaLIFT approach. **Stage-I (Language-Align)**: A language alignment layer is introduced into the LLM to adapt the pre-trained multilingual encoder, thereby enhancing multilingual alignment through code-switched tuning. **Stage-II (Task-Transfer)**: The LLM is fine-tuned on high-quality, English-only instruction data with the language alignment layer frozen, allowing LLM to transfer task-specific capabilities learned from English to low-resource languages.

3.1 Model Architecture

Given a task-specific input x with l_x tokens, we use a multilingual pre-trained model to encode it into a moderately language-agnostic representation X , facilitating more effective cross-lingual understanding and transfer:

$$X = \text{Encoder}(x), \quad (1)$$

where $\text{Encoder}(\cdot)$ is the pre-trained multilingual encoder, and $X \in \mathbb{R}^{l_x \times d_1}$ is the final hidden state.

We introduce the language alignment layer, $\text{Alignment}(\cdot)$, to bridge the gap between the multilingual encoder and the LLM representation, implemented as a multi-layer perceptron (MLP). This layer maps X to the input embedding space of the LLM:

$$\hat{X} = \text{Alignment}(X), \quad (2)$$

where $\hat{X} \in \mathbb{R}^{l_x \times d_2}$ is the alignment representation of X on the input embedding space of the LLM.

To leverage the built-in capabilities of the LLM for instruction following, we embed the instruction context q with l_q tokens using the embedding layer $\text{Embedding}(\cdot)$:

$$Q = \text{Embedding}(q), \quad (3)$$

where $Q \in \mathbb{R}^{l_q \times d_2}$ presents the instruction context q in the input embedding space of the LLM. Then, we concatenate the instruction context representation with the multilingual input representation:

$$(Q, \hat{X}) = [\langle \text{bos} \rangle; Q; \langle \text{enc_start} \rangle; \hat{X}; \langle \text{enc_end} \rangle], \quad (4)$$

where $\langle \text{bos} \rangle \in \mathbb{R}^{d_2}$ is the start token representation and $\langle \text{enc_start} \rangle$ and $\langle \text{enc_end} \rangle$ are trainable boundary parameters of \hat{X} . Finally, (Q, \hat{X}) is fed into the LLM to generate the response.

3.2 Language Alignment

In this stage, we enhance the model's multilingual alignment through code-switched tuning. This involves two key parts: building multilingual alignment lexicons and enhancing alignment via code-switched translation instruction tuning.

Building Multilingual Alignment Lexicons.

Given the scarcity of data for low-resource languages, we adopt an unsupervised word translation method MUSE (Lample et al., 2018) to construct the multilingual alignment lexicons without parallel data. Specifically, we extract a set of unique words $\mathcal{W} = \{w\}_i^N$ from English monolingual corpus \mathcal{D} , where N is the number of

unique words, excluding named entities and stop words. We collect the translations for each word $w \in \mathcal{W}$ by retrieving the top nearest neighbor by projecting the word embeddings into the target space. Further implementation details are provided in Appendix A.1.

Enhancing Alignment with Code-Switched Translation Instruction Tuning. We leverage the built multilingual alignment lexicons to generate code-switched inputs. Formally, this tuning aims to predict the English translation response tokens y given the instruction context q and code-switched input tokens x_l in language l . The input embeddings for this task are represented as (Q, \hat{X}_l) from Equation (4). The optimization objective is outlined as follows:

$$-\arg \min_{\theta} \sum_{l \in \mathcal{L}} \sum_{i=1}^T \log p_{\theta, \sigma, \phi}(y_i | (Q, \hat{X}_l), y_{<i}), \quad (5)$$

where T represents the sequence length of the target output y , $y_{<i}$ denotes the tokens preceding the i -th token, and y_i is the i -th token of y . The trainable parameters include θ , corresponding to the language alignment layer and the boundary tokens. σ and ϕ represent the parameters of the multilingual encoder and LLM, respectively, which remain frozen during training to mitigate the risk of catastrophic forgetting. Finally, \mathcal{L} denotes the set of target languages.

3.3 Task Transfer

After establishing language alignment in the first stage, we fine-tune LLM using high-quality, English-only task-specific instruction data while freezing the language alignment layer. This approach preserves the multilingual alignment and cross-lingual transferability capabilities acquired in the first stage, unlocking the potential of LLM to transfer strong capabilities learned from English to low-resource language tasks. Specifically, we fine-tune the parameters of the LLM to minimize the language modeling negative log-likelihood of the response y , given the combined instruction context and task-specific input (Q, \hat{X}_{en}) :

$$-\arg \min_{\phi} \sum_{i=1}^T \log p_{\bar{\theta}, \sigma, \phi}(y_i | (Q, \hat{X}_{en}), y_{<i}), \quad (6)$$

where ϕ represents the trainable parameters of the LLM, and $\bar{\theta}$ is initialized from the checkpoint of θ

trained at the first stage. The multilingual encoder parameters σ and $\bar{\theta}$ are kept frozen throughout the training.

4 The MMWP Benchmark

Existing multilingual mathematical reasoning benchmarks (Shi et al., 2023; Chen et al., 2023) predominantly focus on 7 high-resource languages and 3 low-resource languages, leaving a significant gap in coverage for low-resource languages. This imbalance introduces evaluation bias, as models are optimized for high-resource languages while their performance on low-resource languages remains underexplored, hindering the comprehensive development of multilingual models. To fill this gap, we build a new Multilingual Math World Problem (MMWP) benchmark and describe its collection process in this section.

Source data. We use AsDiV (Miao et al., 2020) and MAWPS (Koncel-Kedziorski et al., 2016) as base datasets. From the AsDiv official test set, we randomly select 500 examples and another 500 from MAWPS, where all the problems require multiple steps to solve, as described by Miao et al. (2020). We filter out duplicates and problems with non-numeric answers, resulting in 811 examples.

Target Language Selection. We select a typologically diverse set of 48 languages, spanning 12 language families and 12 writing systems, with a range of resource levels—low, medium, and high. This includes 21 low-resource languages², 17 medium-resource languages³, and 10 high-resource languages⁴. Additional selection details are provided in Appendix B.1.

Translation process. To ensure the high quality of translations, we employed Google Translation System to translate the selected English questions into 47 other languages. Moreover, five

²Afrikaans (AF), Arabic (AR), Belarusian (BE), Bengali (BN), Basque (EU), Gujarati (GU), Hausa (HA), Hindi (HI), Armenian (HY), Icelandic (IS), Kannada (KN), Luxembourgish (LB), Macedonian (MK), Malayalam (ML), Marathi (MR), Nepali (NE), Slovak (SK), Swahili (SW), Tamil (TA), Telugu (TE), Thai (TH)

³Bulgarian (BG), Catalan (CA), Czech (CS), Danish (DA), Finnish (FI), Croatian (HR), Hungarian (HU), Indonesian (ID), Korean (KO), Norwegian Bokmål (NB), Polish (PL), Portuguese (PT), Romanian (RO), Slovenian (SL), Serbian (SR), Ukrainian (UK), Vietnamese (VI)

⁴German (DE), English (EN), Spanish (ES), French (FR), Italian (IT), Japanese (JA), Dutch (NL), Russian (RU), Swedish (SV), Chinese (ZH)

annotators then post-edit and calibrate the translation, followed by quality estimation through both human and automatic evaluation, as detailed in Appendix B.2.

5 Experiments

5.1 Datasets

Evaluation Dataset. We use the MMWP and latest multilingual benchmarks, **MGSM** (Shi et al., 2023) and **MSVAMP** (Chen et al., 2023) to evaluate the performance of LLMs in multilingual mathematical reasoning. We evaluate in zero-shot chain-of-thought reasoning (Wei et al., 2022) setting. To further assess the task generalization of LinguaLIFT, we incorporate several challenging multilingual datasets, including **X-CSQA** (Lin et al., 2021) for commonsense reasoning and **XNLI** (Conneau et al., 2018) for natural language inference.

Training Dataset. We utilize English-only instruction data following prior work (Lu et al., 2024; Huang et al., 2024), which include **MetaMathQA** (Yu et al., 2024) for mathematical reasoning, **MultiNLI** (Williams et al., 2018) for natural language inference, and a set of unified commonsense reasoning tasks, comprising the **X-CSQA**, **OpenBookQA** (Mihaylov et al., 2018), **ARC** (Clark et al., 2018), and **QASC** (Khot et al., 2020) datasets. Additionally, to further explore the potential of the proposed method, we incorporate the recent advanced mathematical reasoning instruction dataset, **OpenMathInstruct-2** (Toshniwal et al., 2024). Statistics of the datasets involved are presented in Table 5, and the prompts for each task are given in Appendix A.3.

5.2 Baselines

We consider three categories of baselines: (1) **Mono-SFT** (Luo et al., 2023; Yue et al., 2024; Yu et al., 2024; Toshniwal et al., 2024; Zhu et al., 2024b), a vanilla method that fine-tunes the model on English task datasets. (2) **Multi-SFT** (Chen et al., 2023; Zhu et al., 2024b; She et al., 2024), a translation-based method that fine-tunes the model on multilingual task datasets. (3) **Leveraging External Tools or Models** (Shi et al., 2023; Yoon et al., 2024; Huang et al., 2024), a method that leverages external translation systems or models with existing powerful LLMs. More training details are presented in Appendix A.4.

LLaMA-2-7B as base model	LR.	MR.	HR.	Avg.
BASELINE				
<i>Mono-SFT</i>				
MAmmoTH [†]	6.36	20.4	26.1	15.5
WizardMath [†]	10.1	28.0	32.2	21.0
MetaMath [†]	13.6	37.2	41.8	27.8
OpenMath2 [‡]	19.3	55.5	62.7	41.1
<i>Multi-SFT</i>				
MathOctopus-Parallel [†]	11.0	20.8	25.2	17.4
MathOctopus-MAPO-DPO [†]	18.6	29.8	33.1	25.6
MetaMathOctopus-MAPO-DPO [†]	16.4	36.7	43.9	29.3
QAlign-MetaMathQA [†]	17.1	39.0	44.5	30.6
<i>Leveraging External Tools or Models</i>				
Translate-En [‡] (MetaMath)	27.6	36.4	40.6	33.4
LangBridge [†] (MetaMath)	33.4	36.8	39.4	35.9
MindMerger-Soft [†] (MetaMath)	36.6	40.7	43.2	39.4
(Add Low-Resource) [‡]	36.7	39.8	41.5	38.8
(Low-Resource Retrain) [‡]	33.4	36.8	39.4	35.9
Translate-En [‡] (OpenMath2)	43.1	54.7	61.5	51.0
LangBridge [‡] (OpenMath2)	47.9	53.8	59.2	52.3
MindMerger-Soft [‡] (OpenMath2)	49.5	54.8	61.2	53.8
OUR METHODS				
LinguaLIFT (MetaMath)	41.2	44.3	45.5	43.2
LinguaLIFT (OpenMath2)	55.4	61.2	62.5	58.9

Table 1: Results on the MMWP benchmark. "LR." denotes the average performance for low-resource languages, "MR." for medium-resource languages, and "HR." for high-resource languages. The bold text highlights the highest score among systems. The dagger symbol (†) indicates results obtained using officially released models, while the double dagger symbol (‡) denotes results from our local implementation. The LLM and extra data used in the third baseline category are indicated in brackets.

5.3 Experimental Results

LinguaLIFT demonstrates significant improvements across low-resource languages on MMWP. Experimental results on the MMWP test set, grouped by language resource levels, are presented in Table 1. We elaborate on six key observations from the results: (1) Mono-SFT models exhibit substantial performance drops in low-resource languages and slight drops in medium-resource languages. (2) Multi-SFT models, while outperforming their monolingual counterparts in low-resource settings, still show a considerable performance gap between low-resource and high-resource languages. (3) Models leveraging external translation systems or pre-trained multilingual models achieve limited generalization to unseen languages, and they still lag behind in performance compared to high-resource languages. (4) MindMerger, retrained using additional open-source parallel data in low-resource languages for a fair comparison, performs well in high-resource languages while hardly improving low-resource tasks, demonstrating its limited applicability in low-

resource scenarios. (5) LinguaLIFT significantly enhances low-resource reasoning performance, outperforming all competitive baselines. (6) Incorporating advanced English reasoning datasets (e.g., OpenMathInstruct-2) further enhances low-resource language reasoning, highlighting the importance of adapting to evolving, high-quality data for improving low-resource language reasoning.

These results highlight the importance of comprehensively evaluating multilingual reasoning models across diverse languages and further validate LinguaLIFT’s effectiveness in enhancing reasoning performance for low-resource languages.

LLaMA-2-7B as base model	MGSM			MSVAMP		
	LR.	HR.	Avg.	LR.	HR.	Avg.
BASELINE						
<i>Mono-SFT</i>						
MAmmoTH [†]	3.40	32.4	21.9	6.57	40.1	30.1
WizardMath [†]	4.00	37.7	25.5	15.7	48.5	38.7
MetaMath [†]	4.60	51.7	34.6	15.2	61.2	47.4
OpenMath2 [‡]	5.60	60.2	40.4	18.8	70.5	55.0
<i>Multi-SFT</i>						
MathOctopus-Parallel [†]	28.0	42.4	37.2	33.6	43.8	40.8
MathOctopus-MAPO-DPO [†]	30.6	43.4	38.8	52.5	58.8	56.9
MetaMathOctopus-MAPO-DPO [†]	31.0	55.6	46.7	57.8	67.4	64.5
QAlign-MetaMathQA [†]	26.3	55.9	45.2	48.4	61.5	57.6
<i>Leveraging External Tools or Models</i>						
Translate-En-MetaMath [‡]	39.6	55.1	49.4	47.4	52.3	50.9
LangBridge-MetaMath [†]	38.4	51.9	47.0	43.9	54.5	51.4
MindMerger-Soft-MetaMath [†]	53.1	57.9	56.2	52.7	60.6	58.2
Translate-En-OpenMath2 [‡]	41.3	63.8	55.6	51.9	62.2	59.1
LangBridge-OpenMath2 [‡]	42.6	60.8	54.2	47.9	64.4	59.5
MindMerger-Soft-OpenMath2 [‡]	60.5	67.5	65.0	63.4	74.2	70.9
OUR METHODS						
AlignIFT-MetaMath	55.4	58.8	57.6	56.1	60.6	59.3
AlignIFT-OpenMath2	63.8	66.5	65.5	67.2	74.3	72.2

Table 2: Results on MGSM and MSVAMP datasets. "LR" denotes the average multilingual performance for low-resource languages, while "HR" represents the performance for high-resource languages. The bold text highlights the highest score among systems. The dagger symbol (†) indicates results obtained using officially released models, while the double dagger symbol (‡) denotes results from our local implementation.

LinguaLIFT enhances LLM performance in low-resource languages on MGSM and MSVAMP test sets. Experimental results on the other two widely used multilingual reasoning benchmarks MGSM and MSVAMP are presented in Table 2. These findings are generally consistent with those from the MMWP test sets. LinguaLIFT notably outperforms previous competitive baselines in low-resource languages on both MGSM and MSVAMP test sets, highlighting its strong generalization to various out-of-domain reasoning tasks and robust performance across low-resource language mathematical reasoning tasks.

LLaMA-2-7B as base model	XNLI			X-CSQA		
	LR.	HR.	Avg.	LR.	HR.	Avg.
BASELINES						
<i>Mono-SFT(English-only Task Data)</i>						
Mono-SFT*	58.7	80.1	68.7	28.6	58.6	51.3
<i>Multi-SFT(Multilingual Task Data with Query Translation)</i>						
Multi-SFT*	63.6	81.5	71.9	29.4	48.6	43.8
QAlign*	67.1	80.9	73.5	35.5	57.9	52.3
<i>Freezing LLM with External Tools or Models</i>						
LangBridge*	73.9	79.4	76.5	30.9	37.8	36.1
Translate-En*	71.1	79.6	75.1	42.7	55.5	52.3
MindMerger-Soft*	74.4	83.1	78.4	47.9	65.4	61.0
OUR METHODS						
LinguaLIFT	77.6	83.3	80.3	49.6	65.5	61.5

Table 3: Results on XNLI and X-CSQA datasets. "LR" denotes the average performance for low-resource languages, and "HR" denotes high-resource languages. The bold text highlights the highest score among systems. The asterisk symbol (*) indicates results are taken from the published results of Huang et al. (2024).

Versatile Adaptation Across Various Tasks with LinguaLIFT. As shown in Table 3, LinguaLIFT outperforms all baselines on language understanding tasks, such as XNLI and X-CSQA, demonstrating its efficiency in adapting various tasks than previous work, as it only needs English-only instruction data while others need to translate instruction data to target language constantly. These results highlight LinguaLIFT’s versatility in generalizing beyond mathematical reasoning, making it a universal and effective approach for enhancing low-resource language task performance.

LinguaLIFT Generalizes well on High-Resource Tasks. Although LinguaLIFT is designed for low-resource languages, it also boosts performance in high-resource languages. LinguaLIFT outperforms several competitive baselines on mathematical reasoning tasks and multilingual understanding tasks, as shown in Tables 1, 2 and 3. These results highlight LinguaLIFT’s ability to improve both reasoning and comprehension across low- and high-resource languages.

Overall, LinguaLIFT demonstrates its effectiveness in enhancing low-resource language capabilities without relying on multilingual instruction data. It established itself as a universally applicable and strong potential method for LLMs to advance low-resource language performance across various tasks. More experimental results are provided in Appendix C.

6 Analysis

6.1 Two-Stage Training Ablation Studies

Training Stage Ablation As shown in Figure 3a, removing the Language-Align stage leads to a

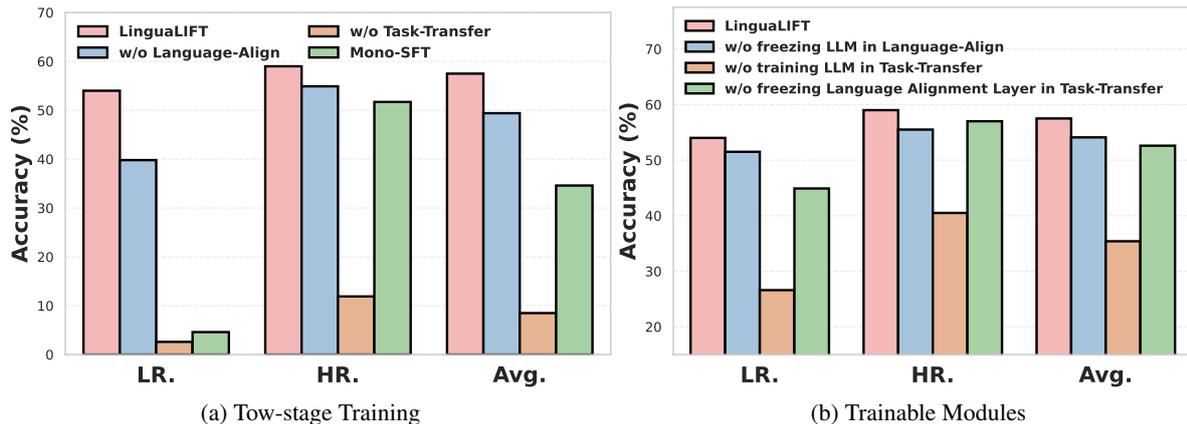


Figure 3: Ablation experiments of two-stage training and trainable modules on MGSM. LR., HR., and Avg. represent the average accuracy across low-resource languages, high-resource languages, and all languages, respectively.

noticeable performance drop in low-resource languages, demonstrating the essential of the Language-Align stage in enhancing low-resource language tasks. Removing the Task-Transfer stage causes substantial performance degradation across all languages. These results suggest the necessity of the two stages. Further details are provided in Appendix D.1.

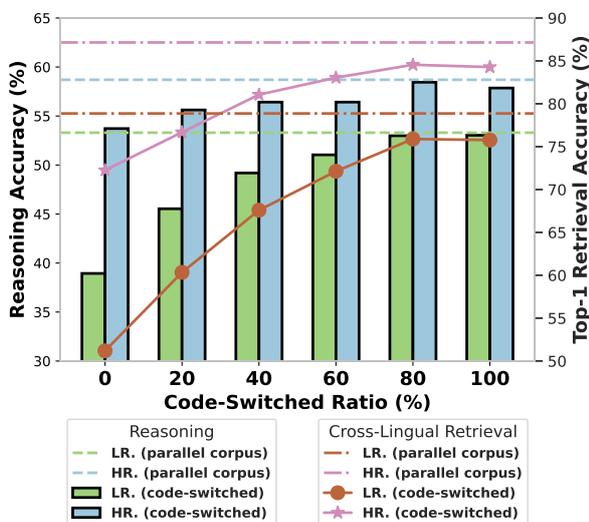


Figure 4: Ablation experiments on the impact of varying code-switch ratios on multilingual alignment and reasoning performance.

Trainable Modules Ablation As illustrated in Figure 3b, training the LLM during the Language-Align stage reduces performance on high-resource language reasoning. Freezing LLM training in the Task-Transfer stage significantly degrades reasoning performance across all languages. Additionally, training the language alignment layer during the Task-Transfer stage harms performance on low-

resource language reasoning. These results indicate the benefit of first training the language alignment layer, followed by LLM fine-tuning. Further details are available in Appendix D.2.

6.2 Analysis of Code-Switched Tuning

Impact of Code-Switched Ratio to Multilingual Alignment and Reasoning Tasks. We report top-1 retrieval accuracy and reasoning accuracy on Tatoeba (Artetxe and Schwenk, 2019) and MGSM (Shi et al., 2023) to indicate language alignment degree and reasoning performance, respectively, as shown in Figure 4. Multilingual alignment improves as the code-switch ratio increases, especially for low-resource languages, and reasoning performance also increases. Notably, at an 80% code-switch ratio, the model’s reasoning performance reaches comparable performance to those trained with parallel corpora.

Impact of the Part-of-Speech in Code-Switched to Reasoning tasks. As shown in Figure 6, among the individual part-of-speech (POS) groups, substituting nouns significantly affects the model’s reasoning ability. Syntactic structures involving subject-verb and prepositional phrase combinations outperform other POS combinations, highlighting the importance of core arguments (subjects and verbs) and their relations (prepositions) for capturing key relationships in reasoning tasks. These structures are sufficient for the model to generalize reasoning across languages. In contrast, adjective-adverb and auxiliary-conjunction combinations perform the worst, suggesting that modifiers are less critical for reasoning tasks. More details are provided in Appendix D.3

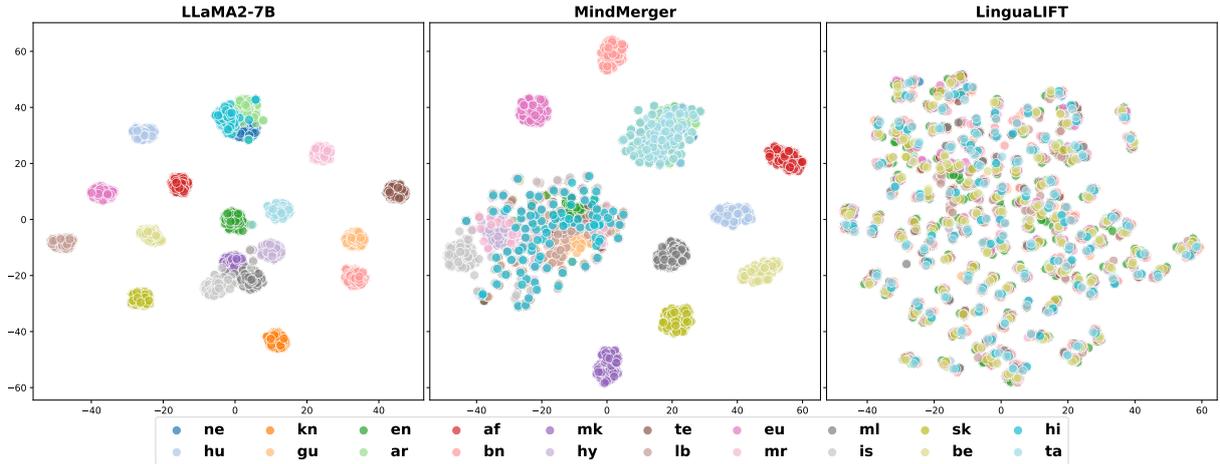


Figure 5: T-SNE visualization in the spaces of the Mono-SFT input embeddings, MindMerger mapping layer outputs, and LinguaLIFT language alignment layer outputs.



Figure 6: The reasoning performance on MGSM group by the combinations of POS: (1) **individual POS categories**: Verb (V), Adposition (ADP), Pronoun (Pron); (2) **syntactic function combinations**: Verb+Adverb (V+Adv), Adjective+Adverb (Adj+Adv), Pronoun+Auxiliary+Conjunction (Pron + Aux + Conj); and (3) **key syntactic structures**: Subject-Verb (N+V) and Prepositional Phrases (N+ADP).

6.3 LinguaLIFT better aligns low-resource languages

For each low-resource language in the MMWP benchmark, we selected 100 texts with equivalent meaning from the Flores-101 dataset (Goyal et al., 2022). We obtained these input mean pooling representations from different methods and visualized them using T-SNE (van der Maaten and Hinton, 2008). As shown in Figure 5, the LLM embeddings of low-resource languages are distinct from English, indicating the challenges of understanding and transferring knowledge to these languages. In the case of MindMerger, some low-resource language representations closely overlap with English, while others remain isolated, demonstrating its limitations in scenarios involving a wide range of low-resource languages. In contrast, LinguaLIFT aligns low-resource language representations more closely with English, enabling better transfer of reasoning capabilities from English instruction

data, thereby resulting in improved reasoning performance.

6.4 Supplementary Experiments

We conducted several supplementary experiments, including the quantitative analysis of the correlation between multilingual alignment and reasoning performance (Appendix E.1), the analysis of language transferability in language families and writing systems (Appendix E.2), the adaptation of different types and scales of LLMs (Appendix E.3, the selection of the language alignment layer and pre-trained multilingual encoder (Appendix E.4, E.5), the incorporation of multilingual instruction data into LinguaLIFT (Appendix E.6), and zero-shot CoT examples in mathematical reasoning tasks (Appendix E.7).

7 Conclusion

This paper introduced LinguaLIFT, a novel two-stage instruction tuning framework that enhances low-resource language tasks without relying on parallel corpora or multilingual instruction data. Additionally, we introduced MMWP, a multilingual benchmark spanning 21 low-resource, 17 medium-resource, and 10 high-resource languages, to comprehensively evaluate multilingual mathematical reasoning tasks. Experiments on the MMWP and other widely used benchmarks demonstrate its effectiveness in advancing low-resource language tasks and further alleviating the performance gap between high-resource and low-resource language tasks in LLMs.

517
518
519
520
521
522
523
524
525
526
527
528
529
530
531

532

533
534
535
536
537
538

539
540
541
542
543

544
545
546
547
548

549
550
551
552

553
554
555
556
557
558
559
560
561

562
563
564
565
566

567
568

Limitations

While our experimental results demonstrate that the proposed two-stage instruction tuning method significantly improves low-resource language reasoning and understanding tasks, it does require a certain level of computational resources. Specifically, the need for a moderately larger pre-trained multilingual encoder and full fine-tuning of the LLM may impose a computational burden. Additionally, we have not explored parameter-efficient fine-tuning (PEFT) methods, so the effectiveness of our approach in PEFT settings remains untested. Future work could focus on developing more efficient and lightweight tuning strategies for LINGUALIFT to reduce the computational limits.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. [Bilingual lexicon induction through unsupervised machine translation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5002–5007, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. [Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond](#). *Transactions of the Association for Computational Linguistics*, 7:597–610.

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*.

Andong Chen, Kehai Chen, Yang Xiang, Xuefeng Bai, Muyun Yang, Tiejun Zhao, et al. 2024a. [Llm-based translation inference with iterative bilingual understanding](#). *arXiv preprint arXiv:2410.12543*.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024b. [DUAL-REFLECT: Enhancing large language models for reflective translation through dual learning feedback mechanisms](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–704, Bangkok, Thailand. Association for Computational Linguistics.

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

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth

Heafield. 2024c. [Monolingual or multilingual instruction tuning: Which makes a better alpaca](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian’s, Malta. Association for Computational Linguistics. 569
570
571
572
573
574

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, Tokyo, Japan. Association for Computational Linguistics. 575
576
577
578
579
580

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. [Scaling instruction-finetuned language models](#). *Journal of Machine Learning Research*, 25(70):1–53. 581
582
583
584
585
586
587
588
589
590
591
592
593

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#). *arXiv preprint arXiv:1803.05457*. 594
595
596
597
598

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics. 599
600
601
602
603
604
605
606
607

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics. 608
609
610
611
612
613
614
615

Zi-Yi Dou, Zhi-Hao Zhou, and Shujian Huang. 2018. [Unsupervised bilingual lexicon induction via latent variable models](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 621–626, Brussels, Belgium. Association for Computational Linguistics. 616
617
618
619
620

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 622
623
624
625
626

627	pages 878–891, Dublin, Ireland. Association for Computational Linguistics.		
628			
629	Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages . In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.		
630			
631			
632			
633			
634			
635			
636			
637			
638			
639			
640			
641	Vedant Gaur and Nikunj Saunshi. 2023. Reasoning in large language models through symbolic math word problems . In <i>Findings of the Association for Computational Linguistics: ACL 2023</i> , pages 5889–5903, Toronto, Canada. Association for Computational Linguistics.		
642			
643			
644			
645			
646			
647	Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation . <i>Transactions of the Association for Computational Linguistics</i> , 10:522–538.		
648			
649			
650			
651			
652			
653			
654	Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation . In <i>Proceedings of the 37th International Conference on Machine Learning</i> , volume 119 of <i>Proceedings of Machine Learning Research</i> , pages 4411–4421. PMLR.		
655			
656			
657			
658			
659			
660			
661			
662	Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting . In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 12365–12394, Singapore. Association for Computational Linguistics.		
663			
664			
665			
666			
667			
668			
669			
670	Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. Mindmerger: Efficiently boosting LLM reasoning in non-english languages . In <i>The Thirty-eighth Annual Conference on Neural Information Processing Systems</i> .		
671			
672			
673			
674			
675	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. 2023. Mistral 7b . <i>arXiv preprint arXiv:2310.06825</i> .		
676			
677			
678			
679			
680	Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning english-centric llms into polyglots: How much multilinguality is needed? <i>arXiv preprint arXiv:2312.12683</i> .		
681			
682			
683			
		Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition . <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , 34(05):8082–8090.	684
			685
			686
			687
			688
		Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. MAWPS: A math word problem repository . In <i>Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 1152–1157, San Diego, California. Association for Computational Linguistics.	689
			690
			691
			692
			693
			694
			695
			696
		Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnab Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2024. Openassistant conversations - democratizing large language model alignment . In <i>Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23</i> , Red Hook, NY, USA. Curran Associates Inc.	697
			698
			699
			700
			701
			702
			703
			704
			705
			706
			707
		Huiyuan Lai and Malvina Nissim. 2024. mCoT: Multilingual instruction tuning for reasoning consistency in language models . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 12012–12026, Bangkok, Thailand. Association for Computational Linguistics.	708
			709
			710
			711
			712
			713
			714
		Guillaume Lample, Alexis Conneau, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data . In <i>International Conference on Learning Representations</i> .	715
			716
			717
			718
		Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation . <i>arXiv preprint arXiv:2305.15011</i> .	719
			720
			721
			722
			723
		Yaoyiran Li, Anna Korhonen, and Ivan Vulić. 2023b. On bilingual lexicon induction with large language models . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 9577–9599, Singapore. Association for Computational Linguistics.	724
			725
			726
			727
			728
			729
		Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning . In <i>Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)</i> , pages 1274–1287, Online. Association for Computational Linguistics.	730
			731
			732
			733
			734
			735
			736
			737
			738
		Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling linguistic horizons	739
			740

741	of LLM by enhancing translation capabilities beyond	798
742	100 languages. In <i>Findings of the Association for</i>	799
743	<i>Computational Linguistics: EMNLP 2024</i> , pages	800
744	10748–10772, Miami, Florida, USA. Association for	801
745	Computational Linguistics.	802
746	Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao,	803
747	Jianguang Lou, Chongyang Tao, Xiubo Geng,	804
748	Qingwei Lin, Shifeng Chen, and Dongmei Zhang.	805
749	2023. Wizardmath: Empowering mathematical	806
750	reasoning for large language models via reinforced	807
751	evol-instruct. <i>arXiv preprint arXiv:2308.09583</i> .	808
752	Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su.	809
753	2020. A diverse corpus for evaluating and developing	810
754	english math word problem solvers. In <i>Proceedings</i>	811
755	<i>of the 58th Annual Meeting of the Association for</i>	812
756	<i>Computational Linguistics</i> , pages 975–984.	813
757	Todor Mihaylov, Peter Clark, Tushar Khot, and	814
758	Ashish Sabharwal. 2018. Can a suit of armor	815
759	conduct electricity? a new dataset for open	816
760	book question answering . In <i>Proceedings of the</i>	817
761	<i>2018 Conference on Empirical Methods in Natural</i>	818
762	<i>Language Processing</i> , pages 2381–2391, Brussels,	819
763	Belgium. Association for Computational Linguistics.	820
764	Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi,	821
765	Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar.	822
766	2024. Gsm-symbolic: Understanding the limitations	823
767	of mathematical reasoning in large language models .	824
768	<i>Preprint</i> , arXiv:2410.05229.	825
769	Niklas Muennighoff, Thomas Wang, Lintang Sutawika,	826
770	Adam Roberts, Stella Biderman, Teven Le Scao,	827
771	M Saiful Bari, Sheng Shen, Zheng Xin Yong,	828
772	Hailey Schoelkopf, Xiangru Tang, Dragomir Radev,	829
773	Alham Fikri Aji, Khalid Almubarak, Samuel Albanie,	830
774	Zaid Alyafeai, Albert Webson, Edward Raff, and	831
775	Colin Raffel. 2023. Crosslingual generalization	832
776	through multitask finetuning . In <i>Proceedings of</i>	833
777	<i>the 61st Annual Meeting of the Association for</i>	834
778	<i>Computational Linguistics (Volume 1: Long Papers)</i> ,	835
779	pages 15991–16111, Toronto, Canada. Association	836
780	for Computational Linguistics.	837
781	Joseph Olive, Caitlin Christianson, and John McCary.	838
782	2011. <i>Handbook of natural language processing</i>	839
783	<i>and machine translation: DARPA global autonomous</i>	840
784	<i>language exploitation</i> . Springer Science & Business	841
785	Media.	842
786	Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida,	843
787	Carroll Wainwright, Pamela Mishkin, Chong Zhang,	844
788	Sandhini Agarwal, Katarina Slama, Alex Ray, John	845
789	Schulman, Jacob Hilton, Fraser Kelton, Luke Miller,	846
790	Maddie Simens, Amanda Askell, Peter Welinder,	847
791	Paul F Christiano, Jan Leike, and Ryan Lowe. 2022.	848
792	Training language models to follow instructions with	849
793	human feedback . In <i>Advances in Neural Information</i>	850
794	<i>Processing Systems</i> , volume 35, pages 27730–27744.	851
795	Curran Associates, Inc.	852
796	Lin Pan, Chung-Wei Hang, Haode Qi, Abhishek Shah,	853
797	Saloni Potdar, and Mo Yu. 2021. Multilingual BERT	854
	post-pretraining alignment . In <i>Proceedings of the</i>	
	<i>2021 Conference of the North American Chapter</i>	
	<i>of the Association for Computational Linguistics:</i>	
	<i>Human Language Technologies</i> , pages 210–219,	
	Online. Association for Computational Linguistics.	
	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-	
	Jing Zhu. 2002. Bleu: a method for automatic	
	evaluation of machine translation . In <i>Proceedings</i>	
	<i>of the 40th Annual Meeting of the Association</i>	
	<i>for Computational Linguistics</i> , pages 311–318,	
	Philadelphia, Pennsylvania, USA. Association for	
	Computational Linguistics.	
	Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and	
	Sebastian Ruder. 2020. MAD-X: An Adapter-	
	Based Framework for Multi-Task Cross-Lingual	
	Transfer . In <i>Proceedings of the 2020 Conference on</i>	
	<i>Empirical Methods in Natural Language Processing</i>	
	<i>(EMNLP)</i> , pages 7654–7673, Online. Association for	
	Computational Linguistics.	
	Maja Popović. 2015. chrF: character n-gram F-score	
	for automatic MT evaluation . In <i>Proceedings of the</i>	
	<i>Tenth Workshop on Statistical Machine Translation</i> ,	
	pages 392–395, Lisbon, Portugal. Association for	
	Computational Linguistics.	
	Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue	
	Huang, and Wanxiang Che. 2023. Cross-lingual	
	prompting: Improving zero-shot chain-of-thought	
	reasoning across languages . In <i>Proceedings of the</i>	
	<i>2023 Conference on Empirical Methods in Natural</i>	
	<i>Language Processing</i> , pages 2695–2709, Singapore.	
	Association for Computational Linguistics.	
	Leonardo Ranaldi, Giulia Pucci, and Andre Fre-	
	itas. 2023. Empowering cross-lingual abilities	
	of instruction-tuned large language models by	
	translation-following demonstrations. <i>arXiv preprint</i>	
	<i>arXiv:2308.14186</i> .	
	Nils Reimers and Iryna Gurevych. 2020. Making	
	monolingual sentence embeddings multilingual using	
	knowledge distillation . In <i>Proceedings of the</i>	
	<i>2020 Conference on Empirical Methods in Natural</i>	
	<i>Language Processing (EMNLP)</i> , pages 4512–4525,	
	Online. Association for Computational Linguistics.	
	Uri Shaham, Jonathan Herzig, Roei Aharoni, Idan	
	Szpektor, Reut Tsarfaty, and Matan Eyal. 2024.	
	Multilingual instruction tuning with just a pinch	
	of multilinguality . In <i>Findings of the Association</i>	
	<i>for Computational Linguistics: ACL 2024</i> , pages	
	2304–2317, Bangkok, Thailand. 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 <i>Proceedings of the 62nd Annual Meeting of the</i>	
	<i>Association for Computational Linguistics (Volume</i>	
	<i>1: Long Papers)</i> , pages 10015–10027, Bangkok,	
	Thailand. Association for Computational Linguistics.	

855	Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang,	Cao, Binbin Xie, et al. 2023. PolyIm: An open	914
856	Suraj Srivats, Soroush Vosoughi, Hyung Won Chung,	source polyglot large language model. <i>arXiv preprint</i>	915
857	Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan	<i>arXiv:2307.06018</i> .	916
858	Das, and Jason Wei. 2023. Language models		
859	are multilingual chain-of-thought reasoners . In	Adina Williams, Nikita Nangia, and Samuel Bowman.	917
860	<i>The Eleventh International Conference on Learning</i>	2018. A broad-coverage challenge corpus for	918
861	<i>Representations</i> .	sentence understanding through inference . In	919
		<i>Proceedings of the 2018 Conference of the</i>	920
862	Shivalika Singh, Freddie Vargus, Daniel D’souza,	<i>North American Chapter of the Association for</i>	921
863	Börje Karlsson, Abinaya Mahendiran, Wei-Yin	<i>Computational Linguistics: Human Language</i>	922
864	Ko, Herumb Shandilya, Jay Patel, Deividas	<i>Technologies, Volume 1 (Long Papers)</i> , pages 1112–	923
865	Mataciunas, Laura O’Mahony, Mike Zhang, Ramith	1122, New Orleans, Louisiana. Association for	924
866	Hettiarachchi, Joseph Wilson, Marina Machado,	Computational Linguistics.	925
867	Luisa Moura, Dominik Krzemiński, Hakimeh		
868	Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib,	Zhihui Xie, Handong Zhao, Tong Yu, and Shuai	926
869	Oshan Mudannayake, Zaid Alyafeai, Vu Chien,	Li. 2022. Discovering low-rank subspaces for	927
870	Sebastian Ruder, Surya Guthikonda, Emad Alghamdi,	language-agnostic multilingual representations . In	928
871	Sebastian Gehrmann, Niklas Muennighoff, Max	<i>Proceedings of the 2022 Conference on Empirical</i>	929
872	Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh	<i>Methods in Natural Language Processing</i> , pages	930
873	Fadaee, and Sara Hooker. 2024. Aya dataset: An	5617–5633, Abu Dhabi, United Arab Emirates.	931
874	open-access collection for multilingual instruction	Association for Computational Linguistics.	932
875	tuning . In <i>Proceedings of the 62nd Annual Meeting</i>		
876	<i>of the Association for Computational Linguistics</i>	Linting Xue, Noah Constant, Adam Roberts, Mihir Kale,	933
877	<i>(Volume 1: Long Papers)</i> , pages 11521–11567,	Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and	934
878	Bangkok, Thailand. Association for Computational	Colin Raffel. 2021. mT5: A massively multilingual	935
879	Linguistics.	pre-trained text-to-text transformer . In <i>Proceedings</i>	936
		<i>of the 2021 Conference of the North American</i>	937
880	Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav	<i>Chapter of the Association for Computational</i>	938
881	Kisacanian, Alexan Ayrapetyan, and Igor Gitman.	<i>Linguistics: Human Language Technologies</i> , pages	939
882	2024. Openmathinstruct-2: Accelerating ai for math	483–498, Online. Association for Computational	940
883	with massive open-source instruction data . <i>arXiv</i>	Linguistics.	941
884	<i>preprint arXiv:2410.01560</i> .		
		Ziyi Yang, Yinfei Yang, Daniel Cer, and Eric Darve.	942
885	Hugo Touvron, Louis Martin, Kevin Stone, Peter	2021. A simple and effective method to eliminate	943
886	Albert, Amjad Almahairi, Yasmine Babaei, Nikolay	the self language bias in multilingual representations .	944
887	Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti	In <i>Proceedings of the 2021 Conference on Empirical</i>	945
888	Bhosale, et al. 2023. Llama 2: Open foundation	<i>Methods in Natural Language Processing</i> , pages	946
889	and fine-tuned chat models . <i>arXiv preprint</i>	5825–5832, Online and Punta Cana, Dominican	947
890	<i>arXiv:2307.09288</i> .	Republic. Association for Computational Linguistics.	948
891	Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-	Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone	949
892	Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel	Kim, Sheikh Shafayat, and Minjoon Seo. 2024.	950
893	Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid,	LangBridge: Multilingual reasoning without mul-	951
894	Freddie Vargus, Phil Blunsom, Shayne Longpre,	tilingual supervision . In <i>Proceedings of the 62nd</i>	952
895	Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer,	<i>Annual Meeting of the Association for Computational</i>	953
896	and Sara Hooker. 2024. Aya model: An instruction	<i>Linguistics (Volume 1: Long Papers)</i> , pages	954
897	finetuned open-access multilingual language model .	7502–7522, Bangkok, Thailand. Association for	955
898	In <i>Proceedings of the 62nd Annual Meeting of the</i>	Computational Linguistics.	956
899	<i>Association for Computational Linguistics (Volume</i>		
900	<i>1: Long Papers)</i> , pages 15894–15939, Bangkok,	Longhui Yu, Weisen Jiang, Han Shi, Jincheng YU,	957
901	Thailand. Association for Computational Linguistics.	Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li,	958
		Adrian Weller, and Weiyang Liu. 2024. Metamath:	959
902	Laurens van der Maaten and Geoffrey Hinton. 2008.	Bootstrap your own mathematical questions for large	960
903	Visualizing data using t-sne . <i>Journal of Machine</i>	language models . In <i>The Twelfth International</i>	961
904	<i>Learning Research</i> , 9(86):2579–2605.	<i>Conference on Learning Representations</i> .	962
905	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	Xiang Yue, Xingwei Qu, Ge Zhang, Yao Fu, Wenhao	963
906	Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le,	Huang, Huan Sun, Yu Su, and Wenhao Chen.	964
907	and Denny Zhou. 2022. Chain-of-thought prompting	2024. MAMMO TH: Building math generalist	965
908	elicits reasoning in large language models . In	models through hybrid instruction tuning . In	966
909	<i>Advances in Neural Information Processing Systems</i> ,	<i>The Twelfth International Conference on Learning</i>	967
910	volume 35, pages 24824–24837. Curran Associates,	<i>Representations</i> .	968
911	Inc.		
		Hongbin Zhang, Kehai Chen, Xuefeng Bai, Yang	969
912	Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei	Xiang, and Min Zhang. 2024a. Paying more	970
913	Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei		

971 attention to source context: [Mitigating unfaithful](#)
972 [translations from large language model](#). In *Findings*
973 [of the Association for Computational Linguistics:](#)
974 [ACL 2024](#), pages 13816–13836, Bangkok, Thailand.
975 Association for Computational Linguistics.

976 Meng Zhang, Yang Liu, Huanbo Luan, and Maosong
977 Sun. 2017. [Adversarial training for unsupervised](#)
978 [bilingual lexicon induction](#). In *Proceedings of*
979 [the 55th Annual Meeting of the Association for](#)
980 [Computational Linguistics \(Volume 1: Long Papers\)](#),
981 pages 1959–1970, Vancouver, Canada. Association
982 for Computational Linguistics.

983 Ruochen Zhang, Samuel Cahyawijaya, Jan Chris-
984 tian Blaise Cruz, Genta Winata, and Alham Fikri
985 Aji. 2023. [Multilingual large language models](#)
986 [are not \(yet\) code-switchers](#). In *Proceedings*
987 [of the 2023 Conference on Empirical Methods](#)
988 [in Natural Language Processing](#), pages 12567–
989 12582, Singapore. Association for Computational
990 Linguistics.

991 Yidan Zhang, Boyi Deng, Yu Wan, Baosong Yang,
992 Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin,
993 Fei Huang, and Jingren Zhou. 2024b. [P-mmeval: A](#)
994 [parallel multilingual multitask benchmark for consis-](#)
995 [tent evaluation of llms](#). *Preprint*, arXiv:2411.09116.

996 Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang,
997 Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen
998 Zhang, Junjie Zhang, Zican Dong, et al. 2023. A
999 survey of large language models. *arXiv preprint*
1000 *arXiv:2303.18223*.

1001 Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan
1002 Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma.
1003 2024. [Llamafactory: Unified efficient fine-tuning of](#)
1004 [100+ language models](#). In *Proceedings of the 62nd*
1005 [Annual Meeting of the Association for Computational](#)
1006 [Linguistics \(Volume 3: System Demonstrations\)](#),
1007 Bangkok, Thailand. Association for Computational
1008 Linguistics.

1009 Wenhao Zhu, Shujian Huang, Fei Yuan, Cheng Chen,
1010 Jiajun Chen, and Alexandra Birch. 2024a. The
1011 power of question translation training in multilingual
1012 reasoning: Broadened scope and deepened insights.
1013 *arXiv preprint arXiv:2405.01345*.

1014 Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She,
1015 Jiajun Chen, and Alexandra Birch. 2024b. [Question](#)
1016 [translation training for better multilingual reasoning](#).
1017 In *Findings of the Association for Computational*
1018 [Linguistics: ACL 2024](#), pages 8411–8423, Bangkok,
1019 Thailand. Association for Computational Linguistics.

1020 Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu,
1021 Shujian Huang, Lingpeng Kong, Jiajun Chen, and
1022 Lei Li. 2024c. [Multilingual machine translation](#)
1023 [with large language models: Empirical results](#)
1024 [and analysis](#). In *Findings of the Association for*
1025 [Computational Linguistics: NAACL 2024](#), pages
1026 2765–2781, Mexico City, Mexico. Association for
1027 Computational Linguistics.

A Implementation Details

A.1 Collection of the Alignment Lexicons

To better construct multilingual alignment lexicons for low-resource language tasks, we leveraged the Unsupervised Bilingual Lexicon Induction (UBLI) (Zhang et al., 2017; Lample et al., 2018; Dou et al., 2018; Artetxe et al., 2019; Li et al., 2023b), which has been proven effective in inducing word translation pairs by aligning independently trained word embeddings in two languages.

Initially, we tokenized the text using the Spacy library’s tokenization function⁵. The resultant word set comprised all words, barring specific named entities, numbers, and date tokens. This step is crucial in ensuring our focus on frequent and pertinent terms that are likely to hold significance in task-specific domains.

Following prior outstanding work MUSE⁶ (Lample et al., 2018), we construct multilingual lexicons using adversarial training to establish a linear mapping between source and target spaces without relying on cross-lingual supervision. This process involves training the model to align the word embeddings of the source and target languages in a shared semantic space. For each word in the source language, we identify the most relevant translations in the target language by projecting the source word embeddings into the target space and retrieving the top nearest neighbor words based on cosine similarity. These translated words form the bilingual lexicons, which are essential for enhancing multilingual understanding. Table 4 provides examples of word translations derived from the multilingual alignment lexicons we constructed, illustrating the effectiveness of this approach.

English	aviation	railway	era	school
Bengali	বিমান চলাচল	রেলপথ	যুগ	স্কুল
Swahili	anga	reli	zama	shule
Telugu	విమానయానం	రైల్వే	యుగం	పాఠశాల
Thai	การบิน	ทางรถไฟ	ยุค	โรงเรียน

Table 4: Word translations examples for English to several low-resource languages.

This unsupervised approach allowed us to obtain diverse translations while preserving the general meaning of the words, which is crucial for cap-

⁵<https://spacy.io/>

⁶<https://github.com/facebookresearch/MUSE>

turing cross-lingual alignment. Moreover, by avoiding context-specific translations, we ensured that the generated translations had broad applicability across multiple tasks and languages. This strategy reduced the reliance on large-scale parallel corpora or pre-existing bilingual dictionaries and provided a robust and scalable solution for building multilingual lexicons in an unsupervised manner.

A.2 Dataset Statistics

Dataset	# Lang	# Samples
Traing Datasets		
MetaMathQA	1	395,000
OpenMathInstruct-2	1	1,000,000
MGSM8KInstruct	10	73,599
MetaMathQA	1	395,000
Multi-NLI	1	392,702
QASC	1	8,134
ARC	1	3,370
OpenBookQA	1	4,957
X-CSQA(En)	1	8,888
Evaluation Datasets		
MMWP	48	38,400
MGSM	11	2,750
MSVAMP	10	10,000
XNLI	15	75,150
X-CSQA	16	16,000

Table 5: Statistics of involved datasets. "#Langs" denotes the number of languages covered by the dataset, and "#Sample" refers to the total number of samples it contains. "Scenario" denotes the usage scenario of the specific dataset discussed in the methodology section.

A.3 Instruction Tuning Prompts

The prompt for code-switched tuning is adapted from Zhang et al. (2023), where the source language, source sentence, and target language are replaced with the relevant translations.

The prompts for mathematical reasoning tasks, natural language inference, and commonsense question answering are modified from Toshniwal et al. (2024) and Lu et al. (2024), with the general instruction being replaced by the specific problems from the training data.

Prompt for Code-Switched Tuning

Translate the following code-switched sentence from {source_lang} to pure {target_lang}:

```
{source_lang}: {source_sentence}
{target_lang}:
```

Prompt for Mathematical Reasoning

Solve the following math problem. Make sure to put the answer (and only the answer) inside $\boxed{\}$.

```
{instruction}
```

Prompt for XNLI

I will give you a premise and a hypothesis. Choose the most appropriate relationship from the following options: Entailment, Neutral, Contradiction.

Premise:

```
{premise}
```

Hypothesis:

```
{hypo}
```

Answer:

Prompt for X-CSQA

Question:

```
{question}
```

Choices:

- A. {choice_A}
- B. {choice_B}
- C. {choice_C}
- D. {choice_D}
- E. {choice_E}

Answer:

A.4 Training Details

We use LlamaFactory⁷ (Zheng et al., 2024) as the training codebase for our experiments. In the first stage, we train only the additional alignment modules (e.g., the language alignment layer and the boundary tokens) for 3 epochs, with a constant learning rate of $6e-4$ and a batch size of 256. In the second stage, we fine-tune all parameters of the LLM for 3 epochs. The learning rate is set to $2e-5$, with a warm-up ratio of 0.05 and a cosine learning

⁷<https://github.com/hiyouga/LLaMA-Factory>

rate scheduler. We also apply a weight decay of $1e-2$ and use a batch size of 128. All experiments are conducted on eight NVIDIA A100 GPUs for a day.

B Details of Constructing the MMWP Benchmark

B.1 Selection of Target Languages and Categorization of Resource-level.

In our study, we employ a unique categorization method for selecting target languages, which deviates somewhat from the conventional definitions based on the abundance or scarcity of linguistic resources. This deviation is primarily due to our focus on LLMs. Rather than adhering to traditional classifications, we opt to categorize languages based on their language distribution in the pre-training corpus used for the LLMs.

Guided by the LLaMA2 technical report (Touvron et al., 2023), we define low-resource languages as those that constitute less than 0.005% of the available multilingual datasets. On the other hand, we categorize languages with a representation percentage between 0.005% and 0.1% as medium-resource languages. This categorization method allows us to incorporate a diverse set of languages into our study, thereby enabling a more effective and wide-ranging assessment of the multilingual reasoning capabilities of our model.

Furthermore, we believe that the comprehensive evaluation of multilingual tasks necessitates the inclusion of languages from various families and scripts. This diversity is crucial in understanding the robustness and versatility of our model, as it allows us to evaluate its performance and generalization capabilities across different linguistic contexts and resource levels. By incorporating languages from various families, scripts, and resource categories, we can ensure more comprehensive coverage in our multilingual settings evaluation.

This unique approach to language selection and categorization provides a more nuanced understanding of language resource levels in LLMs while also ensuring a broader and more diverse evaluation of multilingual tasks for LLMs. It also underscores the importance of considering language families and scripts for the assessment of LLMs, thereby contributing to a more comprehensive and inclusive approach to language model development and evaluation.

B.2 Quality Estimation proves the effectiveness of the proposed MMWP benchmark.

To ensure the quality of the proposed MMWP benchmark, we conducted both human and automatic quality estimation evaluations.

For human evaluation, we employed five annotators, each of whom was responsible for post-calibrating and assessing the quality of the translated dataset. To facilitate this, we utilized GPT-4 for back-translation of the problems into English, a process that allowed the annotators to compare the translated versions with their original counterparts. The evaluators were tasked to scrutinize the translations for grammatical correctness, fluency, and semantic accuracy. Any significant deviations in meaning between the original and translated versions were carefully addressed. The annotators were guided to revise such instances to better align with the original intent, thereby ensuring the preservation of the original linguistic integrity. Moreover, the post-editing process was carried out to maintain native language properties, such as syntactical flexibility and natural phrasing, thus ensuring the fluency and native-like quality of the back-translations. This rigorous human calibration process aimed to certify that the final MMWP dataset was linguistically accurate and culturally appropriate for all represented languages.

The prompt for back-translating the low-resource language texts and assisting the post-editing process are presented below.

GPT-4 Prompt for Back-Translating

You are a translation assistant. Directly translate the mathematical problems from {source_lang} to English without additional explanations.
{source_sentence}

	BLEU \uparrow	chrF \uparrow	TER \downarrow
Average	78.71	86.08	1.09

Table 6: Automatic quality estimation of MMWP using back-translations, showing average results across all languages.

GPT-4 Prompt for Assisting Annotators Post-Editing

You are provided with the following:

1. A low-resource source language text.
2. The back-translated English text (BT).
3. The reference English translation (REF).

Your task is to evaluate the quality of the back-translated English text (BT) based on three criteria: **grammatical correctness**, **fluency**, and **semantic accuracy**. Then, propose **three alternative revisions** to improve the BT. For each revision, explain why it was made and how it improves the translation.

1. **Grammatical Correctness:**
Does the back-translated text adhere to standard English grammar rules (e.g., subject-verb agreement, punctuation, tense consistency)?
2. **Fluency:**
Is the back-translated text natural and smooth? Does it sound like it was written by a native speaker?
3. **Semantic Accuracy:**
Does the back-translated text accurately reflect the meaning of the source language text? Are there any discrepancies in the interpretation of the source?

Provide three revision suggestions for improving the BT:

1. Each revision should aim to enhance either grammatical correctness, fluency, or semantic accuracy.
2. For each suggestion, explain:
 - Why the revision is necessary.
 - Which aspect of translation quality (grammar, fluency, or accuracy) it improves.

1184	Complementing the human evaluation, we	in the main text.	1231
1185	conducted an automatic evaluation to assess the		
1186	MMWP dataset’s translation quality quantitatively.		
1187	In this process, we adopted a back-translation	D.1 Ablation of Two-Stage Training	1232
1188	strategy, translating the MMWP problems into	To demonstrate the necessity of the proposed two-	1233
1189	the target languages and then back into English.	stage instruction tuning approach, we perform	1234
1190	Subsequently, we used widely-accepted automatic	an ablation study on the distinct components	1235
1191	evaluation metrics such as BLEU (Papineni et al.,	of the two-stage process: the Language-Align	1236
1192	2002), chrF (Popović, 2015), and TER (Olive et al.,	stage and the Task-Transfer stage. This study is	1237
1193	2011) to compare the back-translated problems	conducted on the mathematical reasoning datasets	1238
1194	with the original MMWP dataset. These metrics	MGSM and MSVAMP to assess the impact of each	1239
1195	provided a quantitative measure of the overlap	stage’s removal on reasoning performance across	1240
1196	between the original and back-translated versions,	all languages. The detailed results are presented	1241
1197	thereby offering a quantifiable estimation of the	in Tables 14 and 15 for MGSM and MSVAMP,	1242
1198	translation fidelity. The results of this quality	respectively, where “w/o” indicates the absence of	1243
1199	estimation, presented in Table 6, attest to the	the specific stage.	1244
1200	reliable quality of the translations.		
1201	The results from both the human and automatic	When the Language-Align stage was removed,	1245
1202	evaluations indicate that the dataset is properly	we fine-tuned both the language-alignment layer	1246
1203	constructed and adequately reflects the multilingual	and the LLM using only English reasoning data.	1247
1204	nature of the tasks. The human post-editing process	This resulted in a noticeable decline in performance	1248
1205	ensured that the translated problems maintained	in low-resource languages, such as a 10.2% perfor-	1249
1206	their semantic integrity, while the automatic eval-	mance drop in Bengali (Bn), 13.2% performance	1250
1207	uation confirmed that the translations preserved	drop in Thai (Th), 21.2% performance drop in	1251
1208	the necessary linguistic structure for multilingual	Telugu (Te), and 7.6% performance drop in Swahili	1252
1209	reasoning. Overall, the quality assurance mea-	(Sw) on the MGSM dataset and a 14.4% perfor-	1253
1210	sures we implemented guarantee that the MMWP	mance drop in Bengali (Bn), 15.7% performance	1254
1211	benchmark is reliable and effective in evaluating	drop in Swahili (Sw) and 19.8% performance	1255
1212	multilingual mathematical reasoning across diverse	drop in Thai (Th) on the MSVAMP dataset. This decline	1256
1213	languages and resource levels.	highlights the importance of using code-switched	1257
1214	C Complete Experimental Results	multilingual input as a warm-up strategy before	1258
1215	C.1 Evaluation Results on MMWP	directly applying English-only instruction data. By	1259
1216	The complete experimental results on MMWP are	incorporating this stage, LinguaLIFT can better	1260
1217	shown in Table 7, 8 and 9. These tables present	leverage the multilingual model’s representation	1261
1218	Mono-SFT, Multi-SFT, and Leveraging External	space, which contains valuable information from	1262
1219	Tools or Models category comparison baselines.	low-resource languages.	1263
1220	C.2 Evaluation Results on MGSM and	When the Task-Transfer stage was removed, we	1264
1221	MMWP	fine-tuned the language alignment layer using only	1265
1222	The complete experimental results on MGSM and	code-switched translation data without updating	1266
1223	MSVAMP are shown in Table 10, and 11.	the LLM’s parameters. After ablating the Task-	1267
1224	C.3 Evaluation Results on XNLI and	Transfer stage, the models couldn’t complete rea-	1268
1225	X-CSQA	soning tasks across all languages, resulting in an	1269
1226	The complete experimental results on XNLI and	average accuracy of just 8.49% on the MGSM	1270
1227	X-CSQA are shown in Table 10, and 11.	dataset and 9.74% on the MSVAMP dataset. This	1271
1228	D Analysis Experiments Details	demonstrates that task-specific instruction tuning	1272
1229	This section outlines the experimental implementa-	is critical for enabling the LLM to perform task-	1273
1230	tion for the various analysis experiments presented	solving operations effectively.	1274
		These results underscore the complementary	1275
		nature of the two stages: the Language-Align stage	1276
		enhances the cross-lingual transferability within	1277
		models, which allows the models to better transfer	1278
		task-solving capabilities to low-resource languages	1279
		in the Task-Transfer stage.	1280

Resource Level	Language	MAmmoTH-7B	WizardMath-7B	MetaMath-7B	OpenMath2
Low-Resource	af	20.0	25.3	32.4	48.1
	ar	14.8	18.6	23.6	33.1
	be	7.6	12.5	17.9	30.3
	bn	4.8	7.9	9.9	15.7
	eu	2.2	4.2	6.3	6.5
	gu	2.5	3.6	4.8	4.0
	ha	2.8	4.4	5.9	6.4
	hi	7.6	17.0	24.5	34.4
	hy	3.2	3.7	5.8	6.2
	is	6.5	10.1	14.3	24.3
	kn	2.1	3.6	6.4	4.8
	lb	6.8	10.9	14.6	25.7
	mk	15.9	20.1	30.1	45.6
	ml	2.5	4.2	5.2	5.7
	mr	3.0	8.4	9.6	14.9
	ne	2.0	5.8	8.4	12.3
	sk	15.0	23.7	32.3	47.5
sw	3.1	6.5	7.5	8.4	
ta	2.6	6.0	6.3	5.4	
te	2.5	4.9	6.2	4.4	
th	6.0	11.6	13.4	21.5	
Average		6.4	10.1	13.6	19.3
Medium-Resource	bg	23.2	29.1	38.7	59.7
	ca	23.9	31.4	38.8	59.1
	cs	19.7	28.7	37.5	56.7
	da	23.4	30.3	40.3	55.7
	fi	17.3	22.9	32.6	51.1
	hr	19.9	24.8	35.1	54.5
	hu	13.4	24.3	31.0	47.2
	id	22.2	30.0	38.4	57.6
	ko	14.4	25.9	35.0	47.4
	nb	23.2	29.4	37.9	56.5
	pl	21.2	28.0	37.9	57.3
	pt	25.5	36.3	42.9	65.0
	ro	20.4	27.1	39.5	56.1
	sl	18.3	24.5	33.1	51.7
sr	18.4	26.8	36.1	54.0	
uk	20.7	27.6	40.3	58.8	
vi	21.8	28.4	37.0	54.8	
Average		20.4	28.0	37.2	55.5
High-resource	de	25.2	32.3	41.9	62.8
	en	36.6	37.6	42.7	75.0
	es	30.2	33.8	43.4	66.7
	fr	28.7	30.8	43.5	64.4
	it	28.1	34.7	42.9	63.1
	ja	19.7	28.6	40.8	56.4
	nl	23.9	31.2	41.2	58.9
	ru	23.9	30.3	41.6	60.8
	sv	24.2	30.3	42.5	59.7
zh	20.7	32.1	37.9	59.2	
Average		26.1	32.2	41.8	62.7

Table 7: Detailed results(Accuracy) of Mono-SFT baselines on the MMWP benchmark across 48 languages.

Resource Level	Language	MathOctopus-Parallel-7B	MathOctopus-MAPO-DPO-7B	MetaMathOctopus-MAPO-DPO-7B	QAlign-MetaMathQA-7B
Low-Resource	af	19.1	27.5	32.2	37.4
	ar	8.6	17.1	20.4	22.4
	be	14.6	23.2	21.8	18.1
	bn	20.1	28.1	28.4	26.6
	eu	5.6	11.2	9.4	7.5
	gu	6.2	10.9	5.3	6.3
	ha	5.6	10.6	8.8	6.3
	hi	11.1	24.3	14.7	22.2
	hy	6.0	8.4	6.0	6.5
	is	10.5	18.9	15.7	17.4
	kn	5.4	13.9	5.2	5.2
	lb	8.8	15.5	18.0	14.7
	mk	15.9	26.5	32.2	31.3
	ml	5.9	9.9	7.3	6.7
	mr	6.8	15.2	5.4	11.0
	ne	6.5	16.3	9.1	10.2
	sk	16.3	25.8	33.2	31.6
	sw	21.7	26.9	30.7	30.2
ta	4.0	11.0	4.1	6.8	
te	10.2	17.9	4.0	5.1	
th	21.3	31.1	33.8	35.4	
Average		11.0	18.6	16.4	17.1
Medium-Resource	bg	20.6	30.1	36.4	39.2
	ca	22.8	32.9	41.8	41.2
	cs	18.9	28.0	36.3	39.1
	da	22.1	28.7	39.5	40.7
	fi	19.2	27.1	35.5	36.7
	hr	21.5	28.7	33.1	36.4
	hu	19.1	30.0	31.7	35.4
	id	20.8	29.1	36.5	39.5
	ko	18.0	29.4	33.4	38.5
	nb	22.6	28.5	40.3	40.3
	pl	21.0	31.3	38.1	39.2
	pt	24.3	32.4	44.1	43.8
	ro	19.6	30.7	36.4	38.8
	sl	19.2	30.2	32.6	36.3
	sr	18.5	29.4	33.3	37.2
	uk	24.8	31.2	37.7	41.1
vi	20.4	28.9	36.6	39.3	
Average		20.8	29.8	36.7	39.0
High-resource	de	27.7	33.5	41.3	45.1
	en	29.6	36.0	53.5	52.9
	es	25.8	33.7	46.4	46.0
	fr	24.8	32.4	44.1	43.4
	it	24.0	31.6	44.8	43.4
	ja	25.4	33.5	40.6	43.9
	nl	21.7	30.7	39.6	42.2
	ru	26.5	33.8	43.5	44.1
	sv	21.1	30.5	40.9	42.2
	zh	25.8	35.0	44.8	41.7
Average		25.2	33.1	43.9	44.5

Table 8: Detailed results of Multi-SFT on the MMWP benchmark across 48 languages.

Resource Level	Language	Translate-En	Langbridge	MindMerger	AlignIFT-MetaMath
Low-Resource	af	33.0	40.7	40.2	46.4
	ar	34.9	33.4	35.8	42.5
	be	31.2	35.1	37.1	39.6
	bn	27.7	31.9	37.6	39.5
	eu	18.8	30.2	35.4	39.6
	gu	21.6	31.8	35.8	41.0
	ha	19.9	28.0	29.1	34.4
	hi	27.4	35.3	39.4	41.7
	hy	31.5	33.3	34.7	39.7
	is	31.5	34.7	35.6	41.9
	kn	26.7	29.7	33.8	37.4
	lb	29.2	36.0	37.6	43.4
	mk	35.9	40.8	40.7	43.9
	ml	19.3	28.4	37.4	40.7
	mr	25.6	30.5	34.3	37.6
	ne	25.9	31.8	37.9	40.6
	sk	29.2	36.4	38.6	42.5
	sw	27.3	41.2	36.4	47.6
ta	24.4	27.1	35.0	40.5	
te	28.5	29.7	37.1	41.9	
th	30.3	34.9	38.7	41.7	
Average		27.6	33.4	36.6	41.2
Medium-Resource	bg	38.4	40.3	40.6	46.5
	ca	35.3	37.0	41.7	45.4
	cs	37.7	37.4	44.1	47.4
	da	39.4	39.5	42.1	46.0
	fi	34.1	33.4	37.7	41.3
	hr	31.0	33.7	37.4	40.2
	hu	36.1	33.9	37.1	41.4
	id	38.0	36.9	39.8	43.0
	ko	30.7	33.1	39.5	42.6
	nb	38.2	40.0	40.9	48.1
	pl	38.8	38.5	40.6	43.8
	pt	42.4	41.7	42.9	49.1
	ro	38.0	37.5	42.3	43.7
	sl	34.4	34.8	40.4	44.1
	sr	38.1	36.7	42.4	44.8
	uk	36.3	37.9	41.9	43.7
vi	31.6	33.9	39.8	42.0	
Average		36.4	36.8	40.7	44.3
High-resource	de	41.0	40.7	43.9	44.6
	en	42.5	47.1	47.8	49.7
	es	40.6	39.6	43.9	49.0
	fr	40.2	39.3	42.5	45.6
	it	42.6	41.9	42.7	47.7
	ja	39.6	37.0	40.2	40.4
	nl	39.8	41.9	43.5	47.5
	ru	38.9	40.4	42.4	46.5
	sv	38.8	31.0	42.2	41.4
	zh	41.5	35.1	43.3	42.1
Average		40.6	39.4	43.2	45.5

Table 9: Detailed results(Accuracy) of Leveraging External Tools or Models baseline and LinguaLIFT on the MMWP benchmark across 48 languages.

LLaMA-2-7B as base model	Bn	De	En	Es	Fr	Ja	Ru	Sw	Te	Th	Zh	LR.	HR.	Avg.
Baseline														
<i>Mono-SFT</i>														
MAmmoTH [†]	3.20	33.2	47.2	34.4	32.4	23.2	30.4	2.00	1.60	6.80	26.0	3.40	32.4	21.9
WizardMath [†]	4.40	37.2	51.6	43.2	39.2	27.2	36.8	4.40	1.60	5.60	28.8	4.00	37.7	25.5
MetaMath [†]	5.60	56.4	68.0	54.4	56.0	34.8	52.8	5.20	2.00	5.60	39.6	4.60	51.7	34.6
OpenMath2 [‡]	6.80	65.6	79.2	66.0	62.4	42.4	57.6	6.00	2.80	6.80	48.4	5.60	60.2	40.4
<i>Multi-SFT</i>														
MathOctopus-Parallel [†]	31.2	46.0	51.6	42.8	43.6	34.0	40.0	36.4	10.8	33.6	38.8	28.0	42.4	37.2
MathOctopus-MAPO-DPO [†]	33.2	47.2	46.8	43.2	39.6	41.6	40.8	37.2	13.6	38.4	44.8	30.6	43.4	38.8
MetaMathOctopus-MAPO-DPO [†]	35.6	52.4	70.0	58.0	51.2	46.4	56.0	42.4	2.00	44.0	55.2	31.0	55.6	46.7
QAlign-MetaMathQA [†]	29.6	54.0	68.4	57.6	59.2	45.2	58.4	35.6	2.40	37.6	48.8	26.3	55.9	45.2
<i>Leveraging External Tools or Models</i>														
Translate-En-MetaMath [‡]	49.0	59.6	65.6	59.8	56.2	49.0	48.4	37.4	34.6	37.2	47.0	39.6	55.1	49.4
LangBridge-MetaMath [†]	41.2	53.2	62.4	58.0	51.6	39.6	55.2	39.6	28.0	44.8	43.2	38.4	51.9	47.0
MindMerger-Soft-MetaMath [†]	50.4	59.6	67.2	58.4	55.6	50.0	61.6	55.2	52.8	54.2	53.2	53.1	57.9	56.2
Translate-En-OpenMath2 [‡]	52.1	69.8	76.9	71.6	63.4	57.4	54.2	38.4	35.6	39.2	53.2	41.3	63.8	55.6
LangBridge-OpenMath2 [‡]	43.2	63.0	73.8	69.8	58.9	47.6	60.2	41.0	39.8	46.2	52.2	42.6	60.8	54.2
MindMerger-Soft-OpenMath2 [‡]	60.0	69.4	77.8	74.2	63.4	57.8	71.8	61.2	57.2	63.6	58.4	60.5	67.5	65.0
Our Methods														
LinguaLIFT-MetaMath	54.4	62.0	64.8	63.6	56.8	50.0	60.4	55.6	54.0	57.6	54.0	55.4	58.8	57.6
LinguaLIFT-OpenMath2	63.0	69.6	76.0	75.2	62.8	52.0	71.6	64.2	61.2	66.6	58.0	63.8	66.5	65.5

Table 10: Experimental Results on the MGSM Dataset. "LR." "HR." and "Avg." represent the average performance for low-resource languages, high-resource languages, and all languages, respectively. Following prior work (Yoon et al., 2024), we classify Bn, Te, Th, and Sw as low-resource languages, while the remaining languages are categorized as high-resource. The dagger symbol (†) indicates results obtained using officially released models, while the double dagger symbol (‡) denotes results from our local implementation.

D.2 Ablation of Trainable Modules

To further demonstrate the essential design of the two-stage training approach, we conduct an ablation study on the trainable modules during each stage. This study is performed on the mathematical reasoning datasets MGSM and MSVAMP to assess the impact of training each module at the appropriate stage. The detailed results are presented in Tables 16 and 17 for MGSM and MSVAMP, respectively, where "w/o" indicates the absence of a specific operation.

The results reveal that performance on high-resource languages suffers when the LLM is frozen during the Language-Align stage. We attribute this decline to catastrophic forgetting, where the initial training on low-resource languages interferes with the model’s ability to retain its high-resource language capabilities. Specifically, the average performance on high-resource languages decreases by 3.5% on the MGSM dataset and 2.6% on the MSVAMP dataset.

When the LLM training is ablated in the Task-Transfer stage, performance significantly deteriorates across all languages. This indicates that relying solely on the language alignment layer is insufficient for learning high-level reasoning tasks and effectively transferring knowledge to low-resource languages. The average performance

across all languages drops to 35.4% on the MGSM dataset and 37.8% on the MSVAMP dataset.

Furthermore, freezing the language alignment layer during the Task-Transfer stage results in a decline in performance on low-resource languages. This suggests that the alignment learned in the language alignment layer is disrupted during continuous training, as the average performance on low-resource languages drops by 9.1% on the MGSM dataset and 5.3% on the MSVAMP dataset.

These findings emphasize the benefit of training the language alignment layer first, followed by the LLM training, to improve performance on both low-resource and high-resource languages.

D.3 Analysis of Code-Switch Tuning

In our experiment, we aimed to scrutinize the influence of the part-of-speech (POS) of the substitution words on the multilingual reasoning performance in the context of code-switched tuning. This analysis is essential as it provides insights into the role of various word categories and their combinations in the effectiveness of language model tuning and reasoning capabilities.

Our initial step involved an examination of the distribution of part-of-speech (POS) tags in the reasoning queries, the results of which are tabulated in Table 18. This analysis facilitated a preliminary understanding of the prevalence of different word

LLaMA-2-7B as base model	Bn	De	En	Es	Fr	Ja	Ru	Sw	Th	Zh	LR.	HR.	Avg.
BASELINES													
<i>Mono-SFT</i>													
MAmmoTH [†]	6.2	44.1	39.5	45.4	42.3	34.1	38.1	5.1	8.4	37.5	6.57	40.1	30.1
WizardMath [†]	16.4	49.1	56.1	50.5	50.8	45.4	44.8	13.4	17.2	43.1	15.7	48.5	38.7
MetaMath [†]	12.5	63.5	67.2	64.7	64.9	54.2	58.2	16.7	16.5	55.5	15.2	61.2	47.4
OpenMath2 [‡]	19.9	72.7	78.6	72.5	72.9	64.6	68.5	13.1	23.5	63.8	18.8	70.5	55.0
<i>Multi-SFT</i>													
MathOctopus-Parallel [†]	27.8	43.8	46.1	45.4	44.9	41.4	43.8	38.1	34.9	41.4	33.6	43.8	40.8
MathOctopus-MAPO-DPO [†]	48.3	57.6	59.0	59.4	60.0	58.7	58.0	54.6	54.7	59.1	52.5	58.8	56.9
MetaMathOctopus-MAPO-DPO [†]	50.3	67.7	71.5	68.7	67.5	65.4	64.7	61.3	61.7	66.0	57.8	67.4	64.5
QAlign-MetaMathQA [†]	41.9	63.5	65.5	63.0	62.8	58.0	61.5	53.6	49.7	56.2	48.4	61.5	57.6
<i>Leveraging External Tools or Models</i>													
Translate-En-MetaMath [‡]	47.9	43.9	60.6	51.4	50.9	50.4	53.4	43.1	51.3	55.8	47.4	52.3	50.9
LangBridge-MetaMath [†]	39.6	58.8	60.1	56.8	57.9	45.2	53.6	45.8	46.3	49.4	43.9	54.5	51.4
MindMerger-Soft-MetaMath [†]	52.0	61.1	64.5	62.9	60.8	59.0	58.6	54.0	52.1	57.3	52.7	60.6	58.2
Translate-En-OpenMath2 [‡]	56.3	54.1	72.0	59.5	59.3	60.9	64.6	40.2	59.2	64.9	51.9	62.2	59.1
LangBridge-OpenMath2 [‡]	48.0	68.6	71.9	64.8	66.6	56.5	64.4	42.3	53.5	58.3	47.9	64.4	59.5
MindMerger-Soft-OpenMath2 [‡]	61.2	77.2	77.1	73.6	73.5	74.1	70.8	64.8	64.1	72.8	63.4	74.2	70.9
OUR METHODS													
LinguaLIFT-MetaMath	54.3	62.8	65.2	59.9	62.7	57.9	59.7	56.8	57.2	56.0	56.1	60.6	59.3
LinguaLIFT-OpenMath2	65.2	76.4	77.9	74.2	74.5	73.5	71.4	68.6	67.9	72.2	67.2	74.3	72.2

Table 11: Experimental Results on the MSVAMP Dataset. "LR." "HR." and "Avg." represent the average performance for low-resource languages, high-resource languages, and all languages, respectively. Following prior work (Yoon et al., 2024), we classify Bn, Th, and Sw as low-resource languages, while the remaining languages are categorized as high-resource. The dagger symbol (†) indicates results obtained using officially released models, while the double dagger symbol (‡) denotes results from our local implementation.

LLaMA-2-7B as base model	Ar	Bg	Sw	Th	Tr	Ur	El	Hi	Zh	Ru	Vi	De	Fr	Es	En	LR.	HR.	Avg.
BASELINES																		
<i>Mono-SFT</i>																		
Mono-SFT*	60.9	76.7	45.9	55.4	61.9	49.2	63.7	55.7	74.7	77.6	73.7	80.6	82.2	82.2	90.0	58.7	80.1	68.7
<i>Multi-SFT</i>																		
Multi-SFT*	61.7	78.7	56.3	60.1	65.6	57.5	67.0	61.7	79.1	79.7	73.7	82.3	82.9	83.9	88.8	63.6	81.5	71.9
QAlign*	67.0	79.4	65.2	65.2	67.9	62.2	66.5	63.3	76.6	79.2	73.7	80.9	83.1	83.8	89.1	67.1	80.9	73.5
<i>Leveraging External Tools or Models</i>																		
LangBridge*	75.2	79.6	71.7	72.4	74.8	66.9	79.1	71.1	77.4	77.4	78.5	78.8	79.9	80.5	83.4	73.9	79.4	76.5
Translate-En*	68.9	80.8	65.3	69.5	74.5	61.6	79.3	68.7	74.8	76	76.7	80.6	80.4	81.4	87.4	71.1	79.6	75.1
MindMerger-Soft*	76.2	82.4	66.6	71.8	75.7	69.4	78.5	74.7	80.0	80.7	80.3	83.5	83.9	84.4	88.7	74.4	83.1	78.4
OUR METHODS																		
LinguaLIFT	78.4	83.5	75.1	77.0	77.1	72.4	82.2	75.1	80.2	80.5	80.8	83.6	84.0	84.4	89.5	77.6	83.3	80.3

Table 12: Experimental Results on the XNLI Dataset. "LR." "HR." and "Avg." represent the average performance for low-resource languages, high-resource languages, and all languages, respectively. The asterisk symbol (*) indicates results obtained directly from the published results (Huang et al., 2024).

LLaMA-2-7B as base model	Ar	De	En	Es	Fr	Hi	It	Ja	Nl	Pl	Pt	Ru	Sw	Ur	Vi	Zh	LR.	HR.	Avg.
BASELINES																			
<i>Mono-SFT</i>																			
Mono-SFT*	32.3	61.2	76.3	64.0	63.5	32.9	56.0	49.1	57.5	50.6	61.7	56.0	24.2	25.1	50.9	56.5	28.6	58.6	51.3
<i>Multi-SFT</i>																			
Multi-SFT*	28.7	49.1	67.2	54.3	52.1	32.0	50.2	38.7	45.9	45.5	51.2	46.5	27.6	29.2	38.8	43.8	29.4	48.6	43.8
QAlign*	36.3	58.8	75.7	63.1	60.3	37.8	58.3	49.2	56.3	51.3	59.8	56.3	35.1	32.6	50.5	54.8	35.5	57.9	52.3
<i>Leveraging External Tools or Models</i>																			
LangBridge*	30.6	37.4	44.4	38.4	38.2	30.6	39.1	33.9	38.4	39.8	36.3	35.1	31.8	30.5	33.3	39.8	30.9	37.8	36.1
Translate-En*	44.6	57.3	71.3	55.5	57.2	48.4	56.3	47.1	55.0	53.3	54.7	54.4	36.5	41.3	51.8	51.5	42.7	55.5	52.3
MindMerger-Soft*	51.4	67.0	78.1	69.1	68.1	48.4	66.8	53.9	63.8	63.3	67.1	63.7	45.5	46.2	60.6	62.9	47.9	65.4	61.0
OUR METHODS																			
LinguaLIFT	54.1	68.7	78.3	67.9	68.7	50.6	65.7	53.8	63.5	62.9	68.2	63.5	46.7	46.8	60.4	64.1	49.6	65.5	61.5

Table 13: Experimental Results on the X-CSQA Dataset. "LR." "HR." and "Avg." represent the average performance for details of the low-resource, high-resource, and all languages, respectively. The asterisk symbol (*) indicates results obtained directly from the published results (Huang et al., 2024).

classes within the reasoning queries, setting the stage for further investigation into their impact on multilingual reasoning.

Subsequently, we categorized the POS combinations into three distinct groups for a more granular analysis. The first group comprised individual POS

1340
1341
1342

MGSM	Bn	De	En	Es	Fr	Ja	Ru	Sw	Te	Th	Zh	LR.	HR.	Avg.
LinguaLIFT	52.8	62.8	64.8	65.2	55.6	48.4	61.2	53.2	50.4	56.0	54.8	54.0	59.0	57.5
<i>w/o language-align</i>	41.6	56.8	67.2	60.4	58.8	38.4	56.4	45.6	29.2	42.8	46.4	39.8	54.9	49.4
<i>w/o task-transfer</i>	3.20	12.8	14.6	12.4	12.0	9.20	11.2	2.40	1.60	3.60	10.8	2.60	11.9	8.49
Mono-SFT	6.00	58.0	64.8	55.6	55.6	37.6	52.4	4.40	0.00	4.80	40.80	3.80	52.1	34.6

Table 14: Two-stage Training Ablation Experimental Results on the MGSM Dataset. "LR." "HR." and "Avg." represent the average performance for low-resource languages, high-resource languages, and all languages, respectively. Following prior work (Huang et al., 2024), we classify Bn, Te, Th, and Sw as low-resource languages, while the remaining languages are categorized as high-resource.

MSVAMP	Bn	De	En	Es	Fr	Ja	Ru	Sw	Th	Zh	LR.	HR.	Avg.
LinguaLIFT	51.3	62.8	67.2	59.9	62.7	54.3	59.7	56.5	55.2	56.0	54.3	60.4	58.6
<i>w/o language-align</i>	36.9	58.7	63.6	58.8	62.1	53.5	55.3	40.8	35.4	53.1	37.7	57.9	51.8
<i>w/o task-transfer</i>	3.8	10.4	16.5	11.7	15.3	7.7	12.0	2.2	4.1	13.7	3.37	12.5	9.74
Mono-SFT	15.2	59.5	64.9	62.9	61.7	52.4	58.3	15.2	18.3	53.7	16.2	59.1	46.2

Table 15: Two-stage Training Ablation Experimental Results on the MSVAMP Dataset. "LR." "HR." and "Avg." represent the average performance for low-resource languages, high-resource languages, and all languages, respectively. Following prior work (Huang et al., 2024), we classify Bn, Te, Th, and Sw as low-resource languages, while the remaining languages are categorized as high-resource.

MGSM	Bn	De	En	Es	Fr	Ja	Ru	Sw	Te	Th	Zh	LR.	HR.	Avg.
LinguaLIFT	52.8	62.8	64.8	65.2	55.6	48.4	61.2	53.2	50.4	56.0	54.8	54.0	59.0	57.5
<i>w/o freezing LLM in Language-Align</i>	50.6	59.3	62.0	60.1	52.7	45.0	58.2	52.6	48.2	54.5	51.5	51.5	55.5	54.1
<i>w/o training LLM in Task-Transfer</i>	26.4	42.4	47.6	45.6	40.4	30.4	41.2	26.8	24.8	28.4	35.6	26.6	40.5	35.4
<i>w/o freezing Language Alignment Layer in Task-Transfer</i>	46.8	60.2	66.4	59.2	57.8	46.8	58.8	45.2	39.6	48.0	49.6	44.9	57.0	52.6

Table 16: Trainable Modules Ablation Experimental Results on the MGSM Dataset. "LR." "HR." and "Avg." represent the average performance for low-resource languages, high-resource languages, and all languages, respectively. Following prior work (Huang et al., 2024), we classify Bn, Te, Th, and Sw as low-resource languages, while the remaining languages are categorized as high-resource.

MSVAMP	Bn	De	En	Es	Fr	Ja	Ru	Sw	Th	Zh	LR.	HR.	Avg.
LinguaLIFT	51.3	62.8	67.2	59.9	62.7	54.3	59.7	56.5	55.2	56.0	54.3	60.4	58.6
<i>w/o freezing LLM in Language-Align</i>	50.7	59.9	65.4	57.4	55.6	50.1	57.5	55.6	54.9	53.0	53.7	57.0	56.0
<i>w/o training LLM in Task-Transfer</i>	29.3	49.7	51.3	45.6	47.4	28.8	35.8	27.1	24.8	38.3	27.1	42.4	37.8
<i>w/o freezing Language Alignment Layer in Task-Transfer</i>	42.4	60.4	65.6	60.2	60.8	53.3	58.5	53.1	51.6	50.8	49.0	58.8	55.9

Table 17: Trainable Modules Ablation Experimental Results on the MSVAMP Dataset. "LR." "HR." and "Avg." represent the average performance for low-resource languages, high-resource languages, and all languages, respectively. Following prior work (Huang et al., 2024), we classify Bn, Te, Th, and Sw as low-resource languages, while the remaining languages are categorized as high-resource.

Part-of-Speech	#Nums
Noun	84726
Verb	42451
Adposition	37727
Pronoun	25012
Adjective	21790
Conjunction	20764
Auxiliary	16409
Adverb	9150

Table 18: Part-of-speech statistics results from the MGSM English test set.

categories, specifically Verbs, Adpositions, and Pronouns. These categories were selected due to their fundamental role in sentence construction and their potential to affect the meaning and structure of statements significantly.

The second group consisted of syntactic function combinations, including Verb+Adverb, Adjective+Adverb, and Pronoun+Auxiliary+Conjunction combinations. These combinations were chosen based on their syntactic roles and capacity to influence sentence structure and meaning. They are integral to creating complex sentence structures and are often pivotal in conveying nuanced meanings.

The third group focused on key syntactic structures: Subject-Verb (Noun+Verb) and Prepositional Phrases (Adposition+Noun). These structures were selected due to their central role in sentence construction and their potential to encapsulate core semantic information. They form the backbone of many sentence structures and play a crucial role in interpreting a sentence’s meaning.

By analyzing the impact of these POS categories and combinations on the performance of code-switched tuning, we aim to provide a comprehensive understanding of the interplay between syntax and semantics in the context of multilingual reasoning. This analysis will shed light on the importance of different word classes and their combinations in improving the efficacy of multilingual language models.

Further experiments were conducted to investigate the effects of replacing different types of words (e.g., nouns, verbs, prepositions, etc.) on the multilingual reasoning capabilities of LinguaLIFT models. As indicated in Figure 6, we observed that replacing nouns significantly impacted the model’s reasoning performance within the individual POS categories. This finding leads to the hypothesis that reasoning in query sentences may be closely tied to noun-based understanding, possibly due to the

central role nouns play in representing key entities and concepts in the task.

Additionally, we found that syntactic structures involving Subject-Verb combinations and Prepositional Phrases performed the best, surpassing all other POS combinations. This suggests that core arguments (subjects and verbs) and their relational elements (prepositions) are crucial for capturing the key relationships in reasoning tasks.

In contrast, the Adjective+Adverb combination showed the weakest performance, indicating that modifiers are less critical for reasoning tasks. This supports the rationale of previous work (Gaur and Saunshi, 2023; Mirzadeh et al., 2024) that proposed transforming mathematical word problems into symbolic reasoning tasks.

Overall, these findings highlight that, for reasoning tasks, word alignment—particularly involving nouns and verbs—is sufficient to enable the model to generalize reasoning abilities across multiple languages.

E Supplementary Experiments

E.1 Quantitative Analysis of the Correlation between Multilingual Alignment and Reasoning Performance

To evaluate the cross-lingual alignment capabilities of LLMs, we employ the multilingual sentence retrieval benchmark **Tatoeba** (Artetxe and Schwenk, 2019), a widely used dataset for assessing ML-LMs. The dataset includes up to 1,000 sentences per language, along with their English translations. In our study, we focus on a subset of the original benchmark that aligns with the language categories used in the reasoning test sets MGSM, ensuring consistency with our experimental design.

We adopt the evaluation procedure outlined in XTREME (Hu et al., 2020) to evaluate cross-lingual alignment. For each language pair, we compute the cosine similarity between sentences in the source language and their closest counterparts in the target language. Following the methods of Yang et al. (2021) and Xie et al. (2022), we use top-1 retrieval accuracy as a metric to quantify alignment between the two languages. A higher retrieval accuracy score indicates a stronger alignment.

Figure 7 illustrates that the performance of low-resource languages lags significantly behind that of other languages in both mathematical reasoning tasks and the cross-lingual sentence retrieval task.

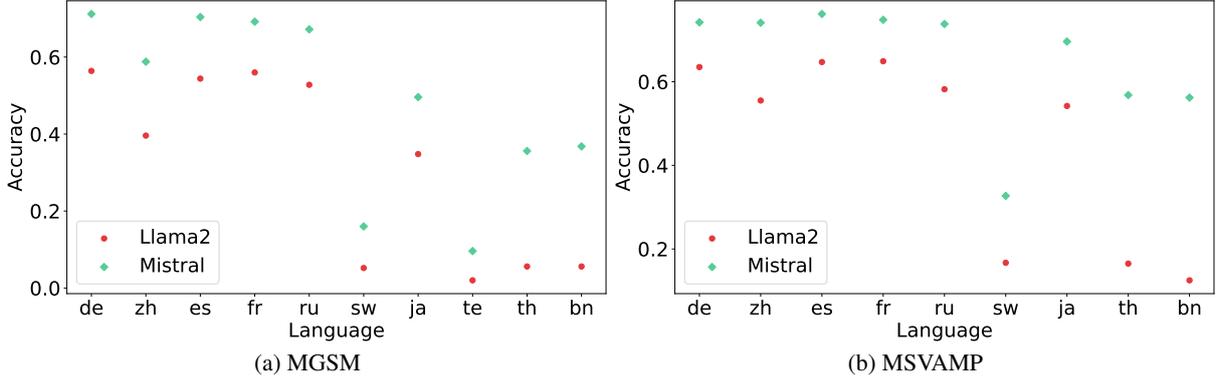


Figure 7: Performances on MGSM and MSVAMP test set, where languages are sorted by decreasing retrieval accuracy. The trend indicates the mathematical reasoning performance worsens along with the decreasing cross-lingual alignment degree.

Test set	MGSM		MSVAMP	
Resource-Level	Low-Resource	High-Resource	Low-Resource	High-Resource
ρ	0.733*	0.778*	0.763*	0.783*

Table 19: Spearman’s rank correlation coefficient (ρ) between retrieval accuracy (%) on Tatoeba and reasoning accuracy on MGSM. An asterisk (*) signifies a statistically significant correlation (p-value < 0.05).

It also shows that a higher alignment degree tends to correlate with improved mathematical reasoning performance. Additionally, we compute the Spearman’s rank correlation between the retrieval accuracy scores and reasoning performance, as presented in Table 19, which reveals a strong correlation between the two. Both the observed trend and the correlation coefficient confirm that the consistency of multilingual mathematical reasoning is closely tied to the cross-lingual alignment degree.

E.2 Language Transferability in Language Families and Writing Systems

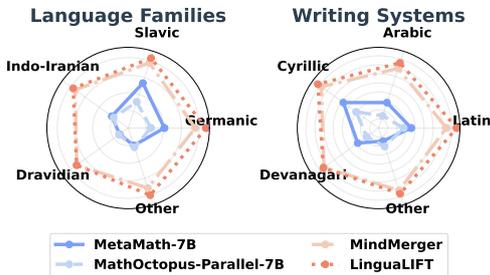


Figure 8: The experimental results grouped by diverse language families and writing systems of different models.

In our analysis, we considered the influence of language families and writing systems on LinguaLIFT’s performance. This approach allowed us

to delve into the nuances of language transferability, particularly in low-resource languages that share linguistic characteristics with English.

We organized the MMWP benchmark results according to language family and writing system, following the categorization proposed by Zhang et al. (2023). As depicted in Figure 8, and detailed results of specific low-resource languages reasoning tasks presented in Table 20, LinguaLIFT demonstrates superior performance over all baseline models across different language families. Notably, it excels in languages from the Indo-European-Indo-Iranian, Indo-European-Germanic, and Dravidian families. The Indo-European-Germanic languages, which share a common lineage with English, register stronger performance, underscoring the potential benefits of shared ancestry in cross-lingual transfer. In contrast, languages that diverge significantly from English in terms of syntax, such as those from the Dravidian family, present a more challenging scenario, yet LinguaLIFT still yields impressive results.

In Figure 8 and in Table 21, we also observed similar trends in the context of writing system transfer. LinguaLIFT exhibits exceptional performance in languages that employ shared scripts, particularly those using Latin-based orthographies. This observation underscores the critical role of script similarity in facilitating the transfer of knowledge across languages. The shared orthographic features

can provide an additional layer of shared structure that aids in the cross-lingual transfer process, enabling more effective learning and translation across related languages.

In summary, our findings underscore the impressive transferability of LinguaLIFT across different low-resource languages from various language families and writing systems. The method significantly enhances cross-lingual transfer, particularly when linguistic factors such as language family and writing system align with English ones. These results highlight the potential of LinguaLIFT as a robust approach for multilingual processing, capable of leveraging shared linguistic characteristics for improved performance, even in low-resource settings.

E.3 Adapting to different types and scales of LLMs

LinguaLIFT can be flexibly adapted to different LLMs. To validate this, we conducted experiments on Mistral-7B (Jiang et al., 2023) and the larger Llama-2-13B (Touvron et al., 2023). As shown in Tables 22 and 23, LinguaLIFT outperforms various baselines, with average accuracy improvements of at least 1.6% and 0.9% for low-resource languages based on Llama-2-13B and Mistral-7B, respectively. These results demonstrate LinguaLIFT’s potential for broader applicability across LLMs.

E.4 Impact of Different Language Alignment Layers

In the main experiment, we utilize two layers of MLP as the language alignment layers to transfer multilingual alignment from the pre-trained multilingual encoder into the LLMs. We also conducted experiments to evaluate the performance of LinguaLIFT when ablating different mapping layers, as shown in Table 24. In contrast to the findings of Yoon et al. (2024), the two layers of MLP used in our main experiment achieved the best performance. In comparison, using a linear layer resulted in lower performance, likely due to its limited capacity to effectively adapt the pre-trained multilingual encoder to LLM, attributed to the smaller number of parameters.

E.5 Impact of Multilingual Encoder Size on Language Alignment

We experimented with various multilingual encoders regarding types and sizes and evaluated the corresponding reasoning performance on the XNLI

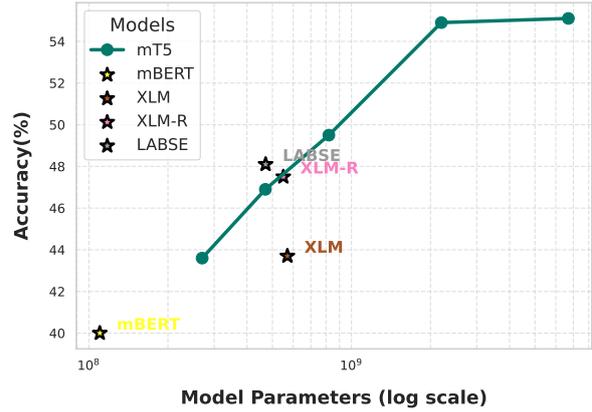


Figure 9: MGSM accuracy(%) of LinguaLIFT models using different kinds and types of multilingual encoders.

dataset. Figure 9 illustrates the XNLI performance across five different sizes of the mT5 encoder: 270M (Small), 470M (Base), 820M (Large), 2.2B (XL), and 6.7B (XXL). Additionally, we evaluate five types of multilingual encoders: mT5, mBERT, XLM, XLM-R, and LaBSE. Our findings indicate that performance improves significantly as the encoder size increases from 270M to 2.2B for the mT5 encoder, with diminishing returns observed as the model scales beyond 2.2B to 6.7B. These results suggest that while enlarging the encoder size leads to better performance, there is a point beyond which further scaling offers limited improvements. In terms of encoder types, we observe that encoders with stronger language alignment capabilities yield better performance. Specifically, more powerful multilingual models like LaBSE, which is fine-tuned with parallel corpus, are more effective at enhancing the capabilities of low-resource languages, demonstrating that greater language alignment leads to better transfer and generalization across diverse languages.

E.6 Incorporating multilingual reasoning data in LinguaLIFT

In some settings, we observed that incorporating multilingual reasoning data during the first stage can further enhance task transfer performance in the second stage. This is evidenced by the improved performance compared to the pure English-only reasoning task training. In the mixed supervision setting, we first fine-tune the language alignment layer during the Language-Align stage using multilingual reasoning instruction data GSM8KINSTRUCT (Chen et al., 2023) and code-switching translation instruction data, followed by fine-tuning LLM with English-only instruction

Language Family	Language	MetaMath-7B	MathOctopus-Parallel-7B	MindMerger	AlignIFT-MetaMath
Indo-European-Germanic	af	32.4	19.1	40.2	46.4
	is	14.3	10.5	35.6	41.9
	lb	14.6	8.75	37.6	43.4
	Average	20.4	12.8	37.8	43.9
Indo-European-Slavic	be	17.9	14.6	37.1	39.6
	mk	30.1	15.9	40.7	42.9
	sk	32.3	16.3	38.6	42.5
	Average	26.8	15.6	38.8	41.7
Indo-European-Indo-Iranian	bn	9.86	20.1	37.6	39.5
	gu	4.81	6.17	35.8	39.0
	hi	24.5	11.1	39.4	41.7
	mr	9.62	6.78	34.3	32.6
	ne	8.38	6.54	37.9	38.6
Average	11.4	10.1	37.0	38.3	
Dravidian	ta	6.29	3.95	35.0	36.5
	te	6.17	10.2	37.1	36.9
	kn	6.41	5.43	33.8	33.4
	ml	5.18	5.92	37.4	36.7
	Average	6.01	6.38	35.8	35.9
Other	ar	23.6	8.63	40.2	42.5
	eu	6.29	5.55	35.4	35.6
	ha	5.92	5.55	29.1	33.4
	hy	5.80	6.04	34.7	39.7
	sw	7.52	21.7	36.4	47.6
	th	13.4	21.3	38.7	41.7
Average	10.4	11.5	35.8	40.1	

Table 20: The detailed experimental results of LinguaLIFT on MMWP benchmark grouped by language family.

Writing Scripts	Language	MetaMath-7B	MathOctopus-Parallel-7B	MingMerger	AlignIFT-MetaMath
Latin	af	32.4	19.1	40.2	46.4
	eu	6.29	5.55	35.4	35.6
	is	14.3	10.5	35.6	41.9
	lb	14.6	8.8	37.6	43.4
	sw	7.52	21.7	36.4	47.6
	sk	32.3	16.3	38.6	42.5
	Average	17.9	13.7	37.3	42.9
Arabic	ar	23.6	8.63	40.2	42.5
	ha	5.92	5.55	29.1	33.4
	Average	14.8	7.09	34.7	38.0
Cyrillic	be	17.8	14.6	37.1	39.6
	mk	30.1	15.9	40.7	42.9
	Average	24.0	15.3	38.9	41.3
Devanagari	hi	24.5	11.1	39.4	41.7
	mr	9.6	6.8	34.3	32.6
	ne	8.4	6.5	37.9	38.6
	Average	14.2	8.1	37.2	37.6
Other	ta	6.29	3.95	35.0	36.5
	te	6.17	10.2	37.1	36.9
	ml	5.18	5.92	37.4	36.7
	bn	9.86	20.1	37.6	39.5
	hy	5.8	6.04	34.7	39.7
	kn	6.41	5.43	33.8	33.4
	gu	4.81	6.17	35.8	39.0
	th	13.4	21.3	38.7	41.7
	Average	7.6	10.8	36.3	38.3

Table 21: The detailed experimental results of LinguaLIFT on MMWP benchmark grouped by writing systems.

Llama-2-13B as base model	Bn	De	En	Es	Fr	Ja	Ru	Sw	Th	Zh	LR.	HR.	Avg.
BASELINE													
<i>Mono-SFT</i>													
MetaMath*	11.6	64.8	67.2	65.2	65.2	42.8	63.6	7.60	6.40	49.2	8.53	59.7	44.4
<i>Multi-SFT</i>													
MathOctopus-Parallel*	35.2	44.4	53.2	48.0	48.4	43.2	47.6	42.8	46.8	48.8	41.6	47.7	45.8
QAlign-MetaMathQA*	38.4	62.0	69.2	67.2	62.4	52.4	64.4	46.0	49.6	59.2	44.7	62.4	57.1
<i>Leveraging External Tools or Models</i>													
Translate-En*	34.8	53.6	70.8	62.4	54.0	44.4	45.6	44.4	54.0	58.0	44.4	55.5	52.2
LangBridge*	39.2	55.2	65.2	60.8	54.8	33.6	58.8	42.0	42.8	42.0	41.3	52.9	49.4
MindMerger-Soft*	55.2	65.2	68.8	69.6	63.6	60.0	68.0	56.4	59.6	60.4	57.1	65.1	62.7
OUR METHODS													
LinguaLIFT	57.6	64.4	67.8	70.4	64.0	59.8	67.2	57.8	60.8	59.4	58.7	64.7	62.9

Table 22: Experimental Results on the MGSM Dataset. "LR." "HR." and "Avg." represent the average performance for low-resource languages, high-resource languages, and all languages, respectively. Following prior work (citation: LLaMA2), we classify Bn, Te, Th, and Sw as low-resource languages, while the remaining languages are categorized as high-resource. The asterisk symbol (*) denotes results taken directly from the published results of [Zhu et al. \(2024b\)](#); [Huang et al. \(2024\)](#).

Mistral-7B as base model	Bn	De	En	Es	Fr	Ja	Ru	Sw	Th	Zh	LR.	HR.	Avg.
BASELINE													
<i>Mono-SFT</i>													
MetaMath*	38.4	70.4	78.0	71.2	70.8	50.8	67.2	16.8	34.8	57.2	30.0	66.5	55.6
<i>Multi-SFT</i>													
MathOctopus-Parallel*	44.0	50.0	58.4	53.2	47.2	48.0	49.6	51.6	48.8	51.6	48.1	51.1	50.2
QAlign-MetaMathQA*	45.6	59.2	65.8	63.6	59.8	49.4	60.2	55.2	51.2	57.2	50.7	59.3	56.7
<i>Leveraging External Tools or Models</i>													
Translate-En*	54.6	50.4	69.7	58.6	56.7	57.2	64.9	47.7	58.7	63.1	53.7	60.1	58.2
LangBridge*	50.0	68.4	65.6	65.6	68.8	58.4	68.4	47.2	60.0	65.6	52.4	65.8	61.8
MindMerger-Soft*	57.6	69.2	79.2	71.6	69.6	57.2	68.4	53.2	59.6	68.8	56.8	69.1	65.4
OUR METHODS													
LinguaLIFT	58.4	67.2	77.6	72.1	70.6	56.4	66.8	58.4	63.2	68.2	60.0	68.4	65.9

Table 23: Experimental Results on the MGSM Dataset. "LR." "HR." and "Avg." represent the average performance for low-resource languages, high-resource languages, and all languages, respectively. Following prior work ([Huang et al., 2024](#)), we classify Bn, Te, Th, and Sw as low-resource languages, while the remaining languages are categorized as high-resource. The asterisk symbol (*) denotes results taken directly from the published results of [Zhu et al. \(2024b\)](#); [Huang et al. \(2024\)](#).

Architecture	# Params	Bn	De	En	Es	Fr	Ja	Ru	Sw	Te	Th	Zh	LR.	HR.	Avg.
Linear	4 M	53.1	61.6	63.9	63.8	56.6	46.0	56.6	49.0	44.4	50.6	52.2	49.3	57.2	54.4
2 layers MLP	10 M	54.4	62.0	64.8	63.6	56.8	50.0	60.4	55.6	54.0	57.6	54.0	55.4	58.8	57.6
3 layers MLP	14 M	53.8	62.8	66.6	56.6	58.8	50.0	59.4	53.8	53.6	58.2	53.4	54.9	58.2	57.0

Table 24: The ablation experiments of the selection of language alignment layer results on the MGSM Dataset. "LR." "HR." and "Avg." represent the average performance for low-resource languages, high-resource languages, and all languages, respectively. Following prior work ([Huang et al., 2024](#)), we classify Bn, Te, Th, and Sw as low-resource languages, while the remaining languages are categorized as high-resource.

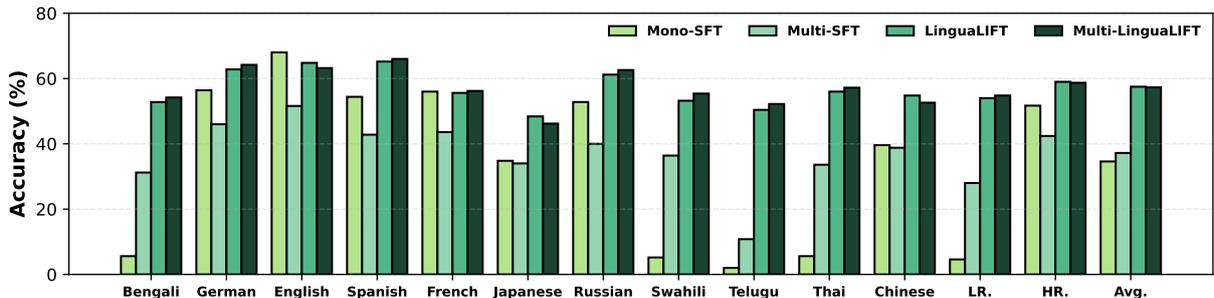


Figure 10: Effects of tuning LLM with mixed supervised data. Generally, incorporating multilingual supervised data into LinguaLIFT can achieve a higher ceiling for low-resource language tasks performance.

1564 data, METAMATHQA. The experimental results on
1565 MGSM are presented in Figure 10. We find that
1566 incorporating additional multilingual supervision
1567 yields an average performance gain of 1.0% on low-
1568 resource reasoning tasks and allows the model to
1569 achieve comparable performance on high-resource
1570 reasoning tasks relative to the vanilla setting.

1571 **E.7 LinguaLIFT CoT Examples**

1572 In this section, we present three examples of CoT
1573 reasoning generated by LinguaLIFT models. These
1574 examples demonstrate the ability of LinguaLIFT to
1575 effectively understand low-resource languages and
1576 generate corresponding mathematical reasoning
1577 processes using the chain-of-thought reasoning
1578 strategy.

1579 Figure 11 illustrates a zero-shot CoT example
1580 from the MGSM Bengali dataset. In this case,
1581 the vanilla Mono-SFT model produces incorrect
1582 reasoning, while the LinguaLIFT model generates
1583 a correct reasoning process, ultimately leading to
1584 the correct answer. Similar trends are observed
1585 in the Thai and Swahili examples, shown in
1586 Figures 12 and 13, where the LinguaLIFT model
1587 corrects the erroneous reasoning produced by the
1588 Mono-SFT model when presented with the same
1589 mathematical problems.

Model Input Query Bn

Solve the following math problem. Make sure to put the answer (and only the answer) inside `\boxed{}`.

প্রশ্ন: টেরি দিনে 2টি ইয়োগার্ট খান। সম্প্রতি 4টি ইয়োগার্ট \$5.00-এর ছাড়যুক্ত মূল্যে বিক্রি হচ্ছে। 30 দিনে তিনি ইয়োগার্টের জন্য কত ডলার ব্যয় করেন?

(Question: Terry eats 2 yogurts a day. 4 yogurts are currently on sale for a discounted price of \$5.00. How many dollars does he spend on yogurt in 30 days?)

Vanilla Response Bn

সম্প্রতি 4টি ইয়োগার্ট \$5.00-এর ছাড়যুক্ত মূল্যে বিক্রি হচ্ছে, তাই এটি $4*5 = \$\langle\langle 4*5=20.00 \rangle\rangle 20.00$ টেরি দিনে 2টি ইয়োগার্ট খান, তাই 30 দিনে তিনি $2*30 = \langle\langle 2*30=60 \rangle\rangle 60$ ইয়োগার্ট খান যদি সম্প্রতি 4টি ইয়োগার্ট \$5.00-এর ছাড়যুক্ত মূল্যে বিক্রি হচ্ছে, তাহলে একটি ইয়োগার্ট খান তার জন্য $5-2 = \$\langle\langle 5-2=3.00 \rangle\rangle 3.00$ যদি টেরি 30 দিনে 60 ইয়োগার্ট খান, তাহলে তিনি $30*3 = \$\langle\langle 30*3=90.00 \rangle\rangle 90.00$ ডলার খরচ করেন।

(Response: Recently 4 yogurts are on sale at a discounted price of \$5.00, so it costs $4*5 = \$\langle\langle 4*5=20.00 \rangle\rangle 20.00$ Terry eats 2 yogurts a day, so in 30 days he eats $2*30 = \langle\langle 2*30=60 \rangle\rangle 60$ yogurts If recently 4 yogurts are on sale at a discounted price of \$5.00, then eating one yogurt costs him $5-2 = \$\langle\langle 5-2=3.00 \rangle\rangle 3.00$ If Terry eats 60 yogurts in 30 days, then he spends $30*3 = \$\langle\langle 30*3=90.00 \rangle\rangle 90.00$ dollars.)

LinguaLIFT Response Bn

Let's think step by step.

Response: Terry eats 2 yogurts per day, so in 30 days, he will eat $2 * 30 = 60$ yogurts.

The 4 yogurts are on sale for \$5.00, so each yogurt costs $\$5.00 / 4 = \1.25 .

Therefore, Terry will spend $60 * \$1.25 = \75.00 on yogurt in 30 days.

`\boxed{75}`

Figure 11: Example of zero-shot CoT generated by MetaMath-7B when prompted with an example from MGSM Bengali. Translation of the input and response is provided in brown text.

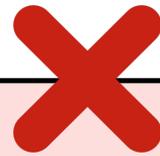
Model Input Query Th

Solve the following math problem. Make sure to put the answer (and only the answer) inside `\boxed{}`.

โจทย์: ซื้อมือถือโทรศัพท์มือถือเครื่องละ \$150 โดยแบ่งผ่อนชำระนาน 3 เดือน โทรศัพท์แต่ละเครื่องจะมีการเรียกเก็บดอกเบี้ย 2% ดังนั้น ใน 3 เดือนนี้ ซื้อมือถือต้องชำระเงินเดือนละกี่ดอลลาร์

(Question: Sheila purchases a \$150 cell phone, to be paid in 3 installments. Each phone is charged 2% interest. How much will Sheila pay each month over the next 3 months?)

Vanilla Response Th



ซื้อมือถือโทรศัพท์มือถือ 3 เครื่อง ดังนั้นต้นทุนรวมคือ $3 * 150$ เหรียญ = 450 เหรียญ อัตราดอกเบี้ยคือ 2% ดังนั้น ดอกเบี้ยรวมคือ $2/100 * 450$ เหรียญ = 9 เหรียญ ดังนั้น ซื้อมือถือต้องจ่ายเงินทั้งหมด 450 เหรียญ + 9 เหรียญ = 459 เหรียญ เนื่องจากเธอต้องผ่อนชำระเป็นเวลา 3 เดือน เธอจึงต้องจ่ายเงิน $459 / 3 = 153$ เหรียญต่อเดือน

(Response: Sheila bought 3 cellphones, so the total cost is $3 * \$150 = \450 . The interest rate is 2%, so the total interest is $2/100 * \$450 = \9 . Therefore, Sheila will have to pay $\$450 + \$9 = \$459$ in total. Since she is paying over 3 months, she will have to pay $\$459 / 3 = \153 per month.)

LinguaLIFT Response Th



Let's think step by step.

Response: Sheila bought a cellphone, so the total cost is $1 * \$150 = \150 .

The interest rate is 2%, so the total interest is $2/100 * \$150 = \3 .

Therefore, Sheila will have to pay $\$150 + \$3 = \$153$ in total.

Since she is paying over 3 months, she will have to pay $\$153 / 3 = \51 per month.

`\boxed{51}`

Figure 12: Example of zero-shot CoT generated by MetaMath-7B when prompted with an example from MGSM Thai. Translation of the input and response is provided in brown text.

Model Input Query Sw

Solve the following math problem. Make sure to put the answer (and only the answer) inside `\boxed{}`.

Swali: Mary mkulima wa bustani mwenye shauku. Jana, alipokea mimea 18 iliyowekwa kwenye nyungu kutoka kwa bustani ya miche anayopendelea zaidi. Tayari ana mimea 2 kwenye nyungu kwenye kila vishubaka 40 vya dirisha lake katika boma lake kubwa la mashambani. Kwa kuwa na ukarimu, ameamua kuwa kesho atapeana mmea 1 ulio kwenye chungu kutoka kwa kila kishubaka kwa marafiki na familia. Mary atasalia na mimea mingapi iliyo kwenye nyungu?

(Question: Question: Mary is an avid gardener. Yesterday, she received 18 potted plants from her favorite nursery. She already has 2 plants in pots on each of her 40 window sills in her large country home. Being generous, she has decided that tomorrow she will give 1 potted plant from each sill to friends and family. How many potted plants will Mary have left?)

Vanilla Response Sw

Mary anapata mimea 18 mpya kutoka kwa bustani ya miche, na tayari ana mimea 2 kwenye nyungu kwa kila kishubaka 40. **Kwa hivyo, ana jumla ya mimea 80 kwenye nyungu (2×40).** Kisha, anaamua kutoa mmea 1 kwa kila kishubaka kwa marafiki na familia kesho. Hivyo, atatoa jumla ya mimea 40 (1 kwa kila kishubaka). **Kwa hiyo, ataachwa na mimea 120 kwenye nyungu.**

(Response: Mary gets 18 new plants from the nursery, and she already has 2 plants in pots for each of the 40 pots. So, she has a total of 80 plants in pots (2×40). Then, she decides to give 1 plant per pot to friends and family tomorrow. So, she will give away a total of 40 plants (1 per pot). So, she will be left with 120 plants in pots.)

LinguaLIFT Response Th

Let's think step by step.

Response: Mary already has 2 plants on each of the 40 planters, so she has a total of $2 \times 40 = 80$ plants.

She also received 18 more plants, so she now has a total of $80 + 18 = 98$ plants.

Mary plans to give away 1 plant from each planter, so she will give away $1 \times 40 = 40$ plants.

After giving away the plants, Mary will have $98 - 40 = 58$ plants left in the planters.

`\boxed{58}`

Figure 13: Example of zero-shot CoT generated by MetaMath-7B when prompted with an example from MGSM Swahili. Translation of the input and response is provided in brown text.