# MERLIN: Multi-Stage Curriculum Alignment for Multilingual Encoder-LLM Integration in Cross-Lingual Reasoning

Anonymous ACL submission

## Abstract

Large language models (LLMs) excel in English but still struggle with complex reasoning in many low-resource languages (LRLs). Existing methods align LLMs with multilingual encoders, such as LangBridge and MindMerger, raising the accuracy for mid and high-resource languages, yet large performance gap remains for LRLs. We present MERLIN, a model-stacking framework that iteratively refines in 2-stages based on a curriculum strategy (from general to specific where general is bilingual bitext and specific is task-specific data) and adapts only a small set of DoRA weights. On the AfriMGSM benchmark MERLIN improves exact-match accuracy by +12.9 pp over MindMerger and outperforms GPT-4o-mini by 15.2 pp. It also yields consistent gains on MGSM and MSVAMP (+0.9 and +2.8 pp), demonstrating effectiveness across both low and high-resource settings.

## 1 Introduction

Large language models (LLMs) have achieved impressive results across a broad range of natural language processing (NLP) tasks (OpenAI et al., 2024; Jiang et al., 2024; Team et al., 2024a). Despite these advances, their performance often deteriorates significantly when dealing with *low-resource languages* (LRLs) (Adelani et al., 2024b; Uemura et al., 2024; Ojo et al., 2024). This shortfall is partly due to the limited online presence of LRLs both in data availability and NLP tools, and heavy focus on high or mid-resource languages during the pretraining of these models (Touvron et al., 2023; Ranathunga et al., 2023; Aryabumi et al., 2024).

To improve reasoning capabilities for LRLs, the most straightforward approach is to adapt LLMs through continual pre-training (Alves et al., 2024; Ng et al., 2025; Buzaaba et al., 2025) or post-training (Ogundepo et al., 2025). However, these methods often depend on access to in-domain data, which is scarce for LRLs due to the lack of annotators, speakers, and linguistic experts. Additionally, fine-tuning LLMs—especially those with billions of parameters—incurs substantial computational costs and energy usage, making these approaches less feasible in resource-constrained settings (Dettmers et al., 2023; Han et al., 2024). LangBridge (Yoon et al., 2024) and MindMerger (Huang et al., 2024) address this challenge by leveraging rich multilingual representations from an external encoder as input to decoder-only LLMs. This design significantly enhances multilingual reasoning capabilities while substantially reducing both data and computational requirements. On the other hand, curriculum learning—a training paradigm in which models are exposed to data of increasing difficulty—shown to improve both convergence speed and final performance (Soviany et al., 2022) and have been applied in recent works to enhance LLMs reasoning ability (Ma et al., 2025).

In this paper, we introduce **M**ultilingual **E**mbedding-Enhanced **R**easoning for **L**anguage **I**ntegration **N**etwork (Figure 1), which is abbreviated to **MERLIN**. It is a two-stage framework that integrates insights from model stacking with question alignment. This design combines data-efficient supervised training with targeted adaptation of the LLM's internal representations, yielding tighter encoder–decoder alignment without the cost of full-model retraining. Inspired by the principles of curriculum learning, in the first stage, we learn a lightweight mapping layer to align a strong multilingual encoder with a frozen LLM through three successive tasks starting with general domain bilingual data, followed by math domain

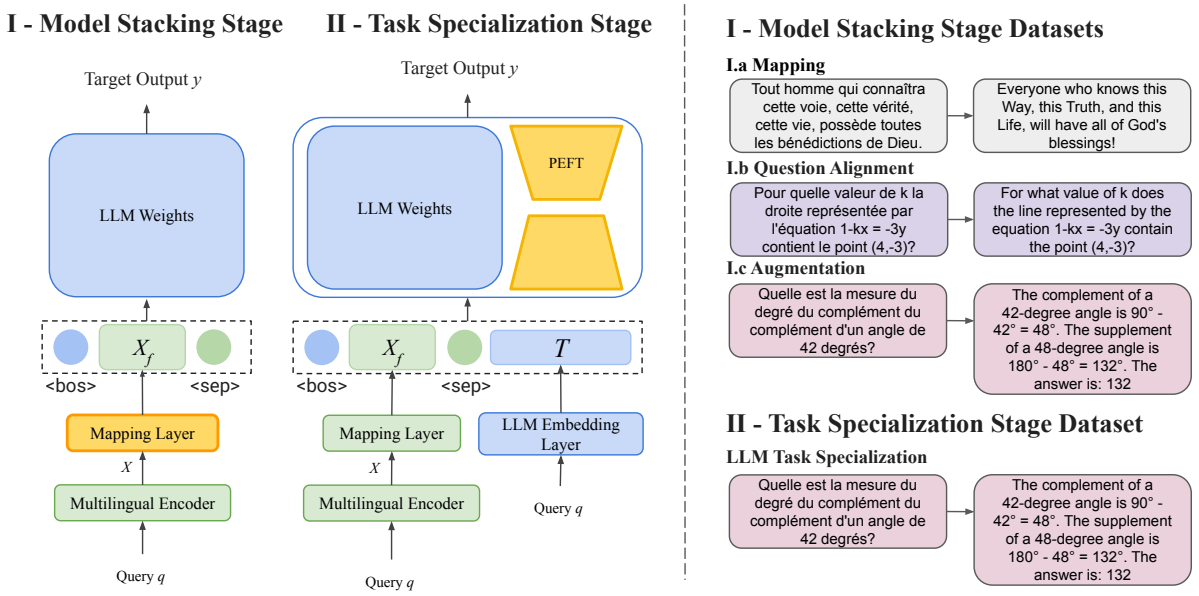


Figure 1: Overview of MERLIN (**M**ultilingual **E**mbedding-Enhanced **R**easoning for **L**anguage **I**ntegration **N**etwork), our two-stage framework for multilingual reasoning. In the Mapping Stage, a lightweight mapping layer is trained sequentially on three datasets—starting with general bilingual translation, followed by query-pair translation, and finally task-specific QA translation—to align the multilingual encoder's outputs with the LLM's embedding space. In the Embedding Enhancement Stage, this mapping "connector" remains frozen while the LLM body is fine-tuned via parameter-efficient PEFT on QA data, thereby strengthening cross-lingual reasoning.

bilingual data, and finally the task data. In the second stage, we fine-tune the LLM with parameter-efficient methods while keeping the mapping layer fixed, enabling the LLM to internalize and better exploit the learned representations during reasoning. We evaluate MERLIN on various math reasoning datasets such as Multilingual Grade School Math Benchmark (MGSM) (Shi et al., 2022), Multilingual Semantic VAriant Math Problems (MSVAMP) (Chen et al., 2024), and AfriMGSM (Adelani et al., 2024b), an extension of MGSM to several low-resource African languages. MERLIN surpasses the strongest baseline, MindMerger, LayAlign, SLAM and QAlign, by more than **+11.0** accuracy points on AfriMGSM and delivers consistent gains on MGSM with **+2.8** performance gain to MindMerger, demonstrating its effectiveness for cross-lingual mathematical reasoning in multilingual settings, including LRLs. To assess generalization beyond mathematical reasoning, we evaluate MERLIN on natural language inference tasks using AfriXNLI (Adelani et al., 2024b), with +1.6 improvement in performance over MindMerger.

## 2 Methodology

Given a frozen multilingual encoder and a lightweight connector into the LLM's embedding space, MERLIN improves cross-lingual reasoning through a two-stage process. The first stage learns to align encoder outputs with the LLM representation space, while the second introduces a small adaptation inside the decoder to internalize this alignment for task solving. The subsections below describe the objectives and data used in each stage.

### 2.1 Stage I: Model Stacking

Stage I is designed to help the decoder learn to interpret encoder representations without modifying its original parameters. This is accomplished through a three-step curriculum that gradually improves cross-lingual alignment, laying the groundwork for the subsequent task specialization stage, where the model's reasoning capabilities are further refined.

**Problem Definition:** Let $q$ be a target-language sentence of length $\ell$. A pretrained multilingual encoder produces $X = \mathrm{Enc}(q) \in \mathbb{R}^{\ell \times d_{\mathrm{enc}}}$. A two-layer perceptron $f_\sigma : \mathbb{R}^{d_{\mathrm{enc}}} \to \mathbb{R}^{d_{\mathrm{llm}}}$, with parameters $\sigma$, projects these states into the LLM space, $X_f = f_\sigma(X)$.

The model stacking stage is divided into three consecutive sub-stages that progressively specialize the connector while the encoder and the LLM remain frozen.

**(a) General bilingual mapping (Map.)** An initial bridge is established with a generic bitext from $L_{\mathrm{tgt}}$, a generic sentence in the target language, to

| Model | Bn | Th | Sw | Ja | Zh | De | Fr | Ru | Es | En | MRL | HRL | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MGSM** | | | | | | | | | | | | | |
| *Open Models* | | | | | | | | | | | | | |
| Metamath 7B | 6.4 | 2.0 | 4.4 | 36.4 | 40.4 | 52.8 | 50.4 | 46.0 | 58.4 | 63.6 | 4.3 | 49.7 | 36.1 |
| Gemma 2 9B | 56.4 | 66.4 | 64.4 | 65.2 | 69.6 | 70.4 | 59.2 | 67.6 | 72.0 | 76.0 | 62.4 | 68.6 | 66.7 |
| *Key Multilingual Approaches–Metamath 7B & mT5-XL* | | | | | | | | | | | | | |
| LangBridge | 42.8 | **50.4** | 43.2 | 40.0 | 45.2 | 56.4 | 50.8 | 52.4 | 58.0 | 63.2 | 45.5 | 52.3 | 50.2 |
| QAlign-LoRA | 32.4 | 39.6 | 40.4 | 44.0 | 48.4 | 54.8 | 56.8 | 52.4 | 59.6 | **68.0** | 37.5 | 54.9 | 49.6 |
| SLAM | 32.0 | 44.0 | 40.0 | 46.0 | 48.4 | 54.0 | 55.2 | 54.8 | 56.8 | 64.8 | 38.7 | 54.3 | 49.6 |
| LayAlign | 1.6 | 3.6 | 8.0 | 23.2 | 23.2 | 37.2 | 37.6 | 34.8 | 38.4 | 43.2 | 4.4 | 33.9 | 25.1 |
| MindMerger | **43.6** | 44.4 | 42.5 | 45.6 | 45.6 | **56.8** | **58.0** | 56.0 | 54.0 | 64.0 | 43.5 | 54.3 | 51.1 |
| MERLIN | 43.2 | 49.2 | **44.8** | **48.4** | **48.8** | 53.2 | 55.6 | **58.0** | **59.2** | 63.6 | **45.7** | **55.3** | **52.4** |
| Δ (MERLIN–MM) | −0.4 | +4.8 | +2.3 | +2.8 | +3.2 | −3.6 | −2.4 | +2.0 | +5.2 | −0.4 | **+2.2** | **+1.0** | **+1.3** |
| *Key Multilingual Approaches–Gemma 2 9B & NLLB-600M-distilled* | | | | | | | | | | | | | |
| QAlign-LoRA | 58.0 | 60.8 | 62.0 | 56.4 | 62.4 | 68.8 | 68.0 | 68.0 | 76.8 | 80.8 | 60.3 | 68.7 | 66.2 |
| SLAM | 63.2 | 61.2 | 52.4 | 62.4 | 64.4 | 68.8 | 53.6 | 70.0 | 69.2 | 69.2 | 58.9 | 65.4 | 63.4 |
| MindMerger | 66.8 | 72.8 | 69.2 | 66.8 | 72.8 | 78.0 | 75.2 | 73.6 | 76.4 | 82.4 | 69.6 | 75.0 | 73.4 |
| **MERLIN** | **74.8** | **76.0** | **77.6** | **71.6** | **72.8** | 73.2 | 74.8 | **78.4** | **77.6** | **85.6** | **76.1** | **76.3** | **76.2** |
| Δ (MERLIN–MM) | +8.0 | +3.2 | +8.4 | +4.8 | 0.0 | −4.8 | −0.4 | +4.8 | +1.2 | +3.2 | **+6.5** | **+1.3** | **+2.8** |
| **MSVAMP** | | | | | | | | | | | | | |
| *Open Models* | | | | | | | | | | | | | |
| Metamath 7B | 11.8 | 13.1 | 11.8 | 40.9 | 48.2 | 56.1 | 58.0 | 54.8 | 59.5 | 63.0 | 12.2 | 54.4 | 41.7 |
| Gemma 2 9B | 63.9 | 72.0 | 71.2 | 78.7 | 81.8 | 78.7 | 79.3 | 78.3 | 80.3 | 80.7 | 69.0 | 79.7 | 76.5 |
| *Key Multilingual Approaches–Metamath 7B & mT5-XL* | | | | | | | | | | | | | |
| LangBridge | 46.8 | 46.3 | 42.1 | 45.5 | 50.4 | 58.1 | 57.0 | 55.8 | 56.9 | 60.6 | 45.1 | 54.9 | 52.0 |
| QAlign-LoRA | 41.7 | 47.7 | 54.8 | 58.0 | 55.7 | 62.8 | 63.2 | 61.1 | 63.3 | 65.3 | 48.1 | 61.3 | 57.4 |
| SLAM | 49.1 | 50.6 | 55.4 | 60.6 | 63.1 | 64.6 | 65.4 | **64.1** | **66.6** | 65.3 | 51.7 | 64.2 | 60.5 |
| LayAlign | 7.3 | 15.4 | 16.9 | 41.7 | 45.9 | 56.4 | 55.3 | 52.9 | 55.0 | 59.3 | 13.2 | 52.4 | 40.6 |
| MindMerger | 51.1 | 47.9 | 53.0 | 54.7 | 53.6 | 58.4 | 60.7 | 57.7 | 60.5 | 64.1 | 50.7 | 58.5 | 56.2 |
| **MERLIN** | **53.1** | **55.6** | **58.9** | **62.0** | **62.3** | **65.5** | **65.7** | 62.2 | 64.9 | **68.2** | **55.9** | **64.4** | **61.8** |
| Δ (MERLIN–MM) | +2.0 | +7.7 | +5.9 | +7.3 | +8.7 | +7.1 | +5.0 | +4.5 | +4.4 | +4.1 | **+5.2** | **+5.9** | **+5.6** |
| *Key Multilingual Approaches–Gemma 2 9B & NLLB-600M-distilled* | | | | | | | | | | | | | |
| QAlign-LoRA | 69.9 | 72.0 | 75.5 | 77.9 | 77.1 | 82.3 | 80.8 | 80.7 | 81.8 | 84.6 | 72.5 | 80.7 | 78.3 |
| SLAM | 60.5 | 70.2 | 68.8 | 74.7 | 74.5 | 74.5 | 74.7 | 73.6 | 75.7 | 72.9 | 66.5 | 74.4 | 72.0 |
| MindMerger | 69.1 | 73.1 | 76.8 | **79.3** | 78.3 | **82.6** | 79.5 | 79.5 | **82.5** | 82.3 | 73.0 | 80.6 | 78.3 |
| **MERLIN** | **72.1** | **76.1** | **77.4** | 78.7 | 79.8 | 81.5 | **81.8** | **79.8** | 81.8 | **83.3** | **75.2** | **81.0** | **79.2** |
| Δ (MERLIN–MM) | +3.0 | +3.0 | +0.6 | −0.6 | +1.5 | −1.1 | +2.3 | +0.3 | −0.7 | +1.0 | **+2.2** | **+0.4** | **+0.9** |

Table 1: Performance on MGSM and MSVAMP. Bn/Th/Sw are mid-resource (MRL); the other seven languages are high-resource (HRL). Sub-headings (e.g. "Metamath-7B & mT5-XL") specify the *decoder base model* and the *multilingual encoder*. LangBridge, LayAlign, MindMerger, and MERLIN use both components; SLAM and QAlign adopt only the listed decoder. Δ rows show MERLIN's absolute gain over MindMerger ($+x$ for improvements, $-x$ for drops).

English (Huang et al., 2024). The decoder receives only the projected encoder states

$$M_\sigma(q) = \big[\langle\mathsf{bos}\rangle;\ X_f;\ \langle\mathsf{sep}\rangle\big],$$

and the connector parameters are obtained from

$$\sigma^{(1)} = \operatorname*{argmin}_\sigma \mathbb{E}_{(\tilde{s},s^\star)}\big[-\log p_{\text{LLM}}\big(s^\star \mid M_\sigma(\tilde{s})\big)\big],$$

where $\tilde{s}$ is the target-language sentence and $s^\star$ its English reference. The loss encourages $M_\sigma(\tilde{s})$ to be a close representation of $s^\star$ in the LLM space.

**(b) Question Alignment (Zhu et al., 2024) (Align.)** The connector is then refined with translated prompts $(\tilde{q}, q^\star)$, where $\tilde{q}$ is a question in $L_{\text{tgt}}$ and $q^\star$ is its English counterpart:

$$\sigma^{(2)} = \operatorname*{argmin}_\sigma \mathbb{E}_{(\tilde{q},q^\star)}\big[-\log p_{\text{LLM}}\big(q^\star \mid M_\sigma(\tilde{q})\big)\big].$$

This step encourages the representations of non-English prompts to align closely with the semantic space utilized by the LLM for English reasoning.

**(c) Task-aware Augmentation (Aug.)** Finally, translated question–answer pairs $(\tilde{q}, a^\star)$ inject task-specific supervision. Here the decoder is allowed to see both the encoder projection and its own token embeddings $T(q) = E_{\text{LLM}}[q]$:

$$(X_f, T)(q) = \big[\langle\mathsf{bos}\rangle;\ X_f;\ \langle\mathsf{sep}\rangle;\ T(q)\big].$$

The objective becomes

$$\sigma^{(3)} = \operatorname{argmin}_\sigma \mathbb{E}_{(\tilde{q},a^\star)}\big[-\log p_{\text{LLM}}\big(a^\star \mid (X_f, T)(\tilde{q})\big)\big],$$

yielding the final connector $\sigma^* = \sigma^{(3)}$.

## 2.2 Stage II: Task Specialization

Stage II internalizes the cross-lingual signal inside the decoder; enhances the integration of multilingual representations within the decoder and plays a critical role in improving performance on complex reasoning tasks. The connector $\sigma^*$, the encoder and all pretrained LLM weights remain fixed; only rank-constrained DoRA (Liu et al., 2024) matrices $\theta$ inside the decoder are activated. For every

English task example $(q, a^\star)$ the input takes the augmented form $(X_f, T)(q)$ defined above. The adaptation parameters are obtained through

$$\theta^* = \operatorname{argmin}_\theta \mathbb{E}_{(q,a^\star)}\big[-\log p_{\mathrm{LLM}_\theta}\big(a^\star \mid (X_f, T)(q)\big)\big].$$

DoRA constrains learning to a negligible fraction of the total parameter count while letting the model gradually absorb how to exploit $X_f$ during generation.

**Inference.** Given a target-language query $q$, the encoder produces $X$, the connector adds the prefix $X_f = f_{\sigma^*}(X)$, and the adapted decoder $\mathrm{LLM}_{\theta^*}$ generates the answer in English from $(X_f, T)(q)$ *zero-shot* with respect to $L_{\mathrm{tgt}}$.

## 3 Experimental Setup

### 3.1 LLMs evaluated

We evaluate two closed models(GPT-4o-mini and GPT-4o) and four open LLMs, including Aya-101 (Üstün et al., 2024), MetaMath 7B, MetaMath-Mistral 7B (Yu et al., 2023), and Gemma 2 9B (Team et al., 2024b). Aya-101 is only evaluated for the AfriXNLI as it is one of the strongest open model baseline for low-resource languages.

### 3.2 Baselines Method Compared

We evaluate our method against five strong baselines: three model merging approaches—LangBridge (Yoon et al., 2024), MindMerger (Huang et al., 2024), LayAlign (Ruan et al., 2025) and two methods leveraging in-domain question translation data, QAlign-LoRA (Zhu et al., 2024) and SLAM (Fan et al., 2025). **LangBridge** injects a frozen multilingual encoder before a decoder-only LLM, replacing token embeddings with linearly projected encoder states (≈3M parameters) learned from 100k English instructions, and inference cost equals that of the LLM plus the multilingual encoder. **QAlign-LoRA** first aligns multilingual prompts via question translation and then fine-tunes lightweight LoRA adapters (<1% parameters) on English tasks. **SLAM** fine-tunes the six lowest feed-forward layers (≈7% parameters) identified as language-sensitive on mixed English and translated data, freezing higher layers and attention weights. **MindMerger** connects a frozen multilingual encoder to the LLM through a two-layer MLP head (≈9M parameters), trained first on translation pairs with English references, then on translated QA pairs while keeping all base weights fixed. **LayAlign** integrates encoder features into multiple LLM layers via a trainable layer-wise aligner, trained in two stages using translation and task data while keeping both encoder and LLM frozen.

For the encoder component in multilingual-integration approaches, we couple both MindMerger and Merlin with the distilled 600M-parameter NLLB encoder (Team et al., 2022). For LangBridge, we follow its original setup: AfriTeva-v2-Large (Oladipo et al., 2023) is used on `AfriMGSM` and `AfriXNLI`, whereas mT5-XL is employed for `MGSM`, because the publicly released LangBridge implementation only supports encoders that subclass the `XXXEncoderModel`.

### 3.3 Languages and Tasks

Our evaluation spans four multilingual benchmarks. **AfriMGSM** (Adelani et al., 2024b) extends MGSM with grade-school mathematics problems translated into 16 African languages, providing fine-grained coverage of more low-resource settings. **AfriXNLI** (Adelani et al., 2024b) offers natural-language inference data in the same 16 languages by translating ten domain splits from the original XNLI. The canonical **MGSM** corpus (Shi et al., 2022) supplies math QA instances across eight mid-resource languages. Finally, **MSVAMP** (Chen et al., 2024) introduces multilingual variants of SVAMP with 1000 algebra-word problems in eleven languages, testing cross-lingual reasoning beyond GSM-style arithmetic.

### 3.4 Experimental Settings

**Training Datasets** **Sub-stage Ia (Mapping).** For every language we sample 9 000 general-domain English–target sentence pairs from the NLLB training corpus (Team et al., 2022). These bitext pairs supply a coarse bilingual anchor for the connector.

**Sub-stage Ib (Question Alignment).** To align non-English prompts with English semantics, we translate 20 000 English questions with `NLLB` 200-3.3B and sample 3 000 translated pairs per language. For the mathematics track (`AfriMGSM` and `MGSM`) the source questions come from `MetaMathQA` (Yu et al., 2024). For the NLI track (`AfriXNLI`) we instead sample premise–hypothesis pairs from translated `MultiNLI` (Williams et al., 2018), concatenate them as the "question", translate, and again retain 3 000 pairs.

| Model | amh | ewe | hau | ibo | kin | lin | lug | orm | sna | sot | swa | twi | wol | xho | yor | zul | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *closed models* | | | | | | | | | | | | | | | | | |
| GPT-4o-mini | 31.6 | 6.0 | 56.0 | 33.6 | 48.0 | 25.6 | 29.2 | 39.2 | 44.8 | 36.8 | 70.8 | 15.6 | 7.6 | 32.4 | 45.6 | 43.6 | 35.4 |
| GPT-4o | 57.6 | 8.8 | 64.8 | 57.6 | 60.4 | 51.2 | 51.6 | 61.2 | 58.4 | 60.8 | 78.8 | 31.2 | 28.0 | 52.4 | 62.0 | 57.2 | 52.6 |
| *open models* | | | | | | | | | | | | | | | | | |
| MetaMath-Mistral 7B | 1.2 | 2.4 | 2.0 | 2.0 | 2.4 | 3.6 | 4.4 | 2.0 | 1.2 | 2.4 | 10.4 | 2.8 | 2.8 | 1.6 | 1.6 | 1.6 | 2.8 |
| Gemma 2 9B | 26.4 | 4.8 | 35.2 | 11.6 | 22.4 | 11.6 | 17.6 | 9.6 | 26.0 | 22.0 | 61.6 | 9.2 | 3.6 | 20.8 | 13.6 | 21.2 | 19.8 |
| *Key multilingual approaches-MetaMath Mistral 7B & AfriTeVa* | | | | | | | | | | | | | | | | | |
| LangBridge | 31.6 | 2.0 | 41.2 | 14.0 | 28.4 | 4.0 | 11.2 | 17.2 | 30.0 | 29.6 | 54.4 | 2.8 | 0.8 | 18.4 | 21.2 | 27.2 | 20.9 |
| QAlign-LoRA | 4.4 | 3.6 | 3.2 | 1.6 | 5.6 | 7.6 | 3.2 | 4.8 | 5.6 | 4.8 | 18.4 | 2.4 | 3.6 | 4.0 | 4.4 | 4.8 | 5.1 |
| SLAM | 1.6 | 2.0 | 1.6 | 1.6 | 1.2 | 2.0 | 2.4 | 1.6 | 1.2 | 4.0 | 2.4 | 4.4 | 2.0 | 2.8 | 2.4 | 1.6 | 2.2 |
| LayAlign | 30.0 | 6.4 | 36.4 | 27.2 | 32.0 | 16.0 | 21.6 | 26.8 | 31.6 | 30.0 | 42.8 | 4.0 | 6.0 | 23.6 | 30.4 | 29.6 | 24.6 |
| MindMerger | 34.4 | 12.0 | 54.4 | 33.6 | 43.2 | 22.0 | 30.0 | 28.8 | 40.4 | 42.8 | **60.0** | **8.4** | 7.2 | 32.0 | 41.2 | 40.8 | 33.2 |
| MERLIN | **45.6** | **12.8** | **60.8** | **37.6** | **48.8** | **26.4** | **33.6** | **39.6** | **44.0** | **44.4** | 58.8 | 7.6 | **9.2** | **38.0** | **43.2** | **41.2** | **37.0** |
| Δ (MERLIN–MM) | +11.2 | +0.8 | +6.4 | +4.0 | +5.6 | +4.4 | +3.6 | +10.8 | +3.6 | +1.6 | −1.2 | −0.8 | +2.0 | +6.0 | +2.0 | +0.4 | **+3.8** |
| *Key multilingual approaches-Gemma2 9B & NLLB-600M distilled* | | | | | | | | | | | | | | | | | |
| QAlign-LoRA | 45.2 | 9.6 | 44.8 | 29.6 | 30.8 | 19.2 | 21.6 | 14.8 | 31.2 | 30.4 | 65.6 | 12.4 | 6.4 | 26.8 | 26.0 | 30.4 | 33.7 |
| SLAM | 51.6 | 26.0 | 43.2 | 33.6 | 41.6 | 37.2 | 33.6 | **45.6** | 40.0 | 33.2 | 57.6 | 21.2 | 18.8 | 30.4 | 41.2 | 34.8 | 39.0 |
| MindMerger | 45.2 | 32.4 | 44.8 | 34.4 | 43.2 | 38.4 | 36.4 | 40.0 | 44.8 | 36.8 | 53.2 | 24.4 | 21.2 | 33.6 | 34.8 | 40.4 | 37.7 |
| MERLIN | **59.2** | **41.2** | **60.8** | **52.0** | **52.0** | **49.6** | **46.8** | 32.0 | **58.8** | **56.0** | **76.8** | **37.6** | **26.0** | **51.2** | **54.0** | **55.6** | **50.6** |
| Δ (MERLIN–MM) | +14.0 | +8.8 | +16.0 | +17.6 | +8.8 | +11.2 | +10.4 | −8.0 | +14.0 | +19.2 | +23.6 | +13.2 | +4.8 | +17.6 | +19.2 | +15.2 | **+12.9** |

Table 2: **AfriMGSM performance across various LLMs and key multilingual approaches**. We report Exact Match scores averaged over the displayed languages. We use the NLLB-600M distilled model as the encoder, and Gemma2-9b-it for LayAlign, MindMerger, MERLIN, and as the base model for QAlign. We use AfriTeVa-v2-Large as the encoder of LangBridge. Note that the released LangBridge source code does not accept NLLB architecture.

| Model | eng | fra | amh | ewe | hau | ibo | kin | lin | lug | orm | sna | sot | swa | twi | wol | xho | yor | zul | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *closed models* | | | | | | | | | | | | | | | | | | | |
| GPT-4o-mini | 86.2 | 80.3 | 52.2 | 37.2 | 64.8 | 57.2 | 56.8 | 31.5 | 51.7 | 61.3 | 63.0 | 56.2 | 63.7 | 47.5 | 43.2 | 64.5 | 54.5 | 62.0 | 54.2 |
| GPT-4o | 89.2 | 82.3 | 71.8 | 45.0 | **75.2** | 68.2 | 68.0 | 32.7 | 69.8 | **71.2** | 71.3 | 71.8 | 71.5 | 55.8 | 52.7 | 72.0 | 64.5 | 67.5 | 64.3 |
| *open models* | | | | | | | | | | | | | | | | | | | |
| Aya-101 | 67.0 | 59.7 | 64.2 | 43.2 | 57.0 | 55.5 | 54.3 | 33.5 | 51.7 | 51.5 | 55.7 | 52.2 | 56.5 | 47.0 | 36.7 | 55.2 | 54.5 | 55.3 | 51.5 |
| Gemma 2 9B | 55.3 | 50.7 | 43.2 | 35.3 | 47.0 | 40.7 | 40.2 | 32.8 | 38.8 | 37.8 | 42.3 | 40.2 | 46.3 | 37.2 | 35.2 | 42.5 | 41.3 | 44.5 | 40.3 |
| *Key Multilingual Approaches-Gemma2 & NLLB 600m Distilled* | | | | | | | | | | | | | | | | | | | |
| QAlign-LoRA | 81.0 | 74.7 | 51.3 | 35.7 | 55.8 | 49.6 | 43.3 | 34.3 | 41.5 | 39.7 | 48.5 | 48 | 60.7 | 43.8 | 35.0 | 51.5 | 47.5 | 49.3 | 49.5 |
| SLAM | 67.0 | 61.8 | 48.2 | 42.7 | 47.2 | 51.3 | 41.7 | 32.2 | 47.5 | 47.7 | 46.5 | 47.2 | 52.8 | 38.8 | 35.7 | 49.8 | 52.2 | 48.3 | 47.7 |
| MindMerger | 80.8 | 76.3 | 69.2 | **60.5** | 70.8 | 68.5 | 66.8 | 34.5 | **70.3** | 54.7 | **75.2** | 71.2 | 71.3 | 63.7 | 59.5 | **73.8** | 68.8 | **74.3** | 67.1 |
| MERLIN | **90.8** | **86.7** | **74.8** | 60.0 | 73.3 | **71.7** | **68.8** | **34.8** | 69.2 | 59.3 | 71.3 | **72.0** | **72.2** | **65.5** | **62.2** | 73.7 | **69.2** | 70.2 | **68.7** |

Table 3: **AfriXNLI performance on various LLMs, MindMerger, and MERLIN**. Metric is Accuracy. Average computed on only African languages. NLLB 600M distilled is used as the encoder.

**Sub-stage Ic (Augmentation).** For each target language we extract a distinct set of 3 000 question–answer pairs from the training split of the task corpus: translated `MetaMathQA` for the math benchmarks (`AfriMGSM`, `MGSM`, and `MSVAMP`) and translated `MultiNLI` for the NLI benchmark (`AfriXNLI`).

**Task-Specialization Stage.** We fine-tune the LLM (via PEFT) on a second sample of 3 000 training instances per language drawn from the same translated sources as above. Using a separate set of examples helps reduce the risk of overlap with Sub-stage Ic and ensures there is no data leakage into the evaluation sets.[1]

## 4 Experimental Results

We assessed MERLIN on both math reasoning and NLI benchmarks, comparing it against open LLMs, recent multilingual integration methods, and the strongest available closed models. For math reasoning, we use two datasets: MSGM and MSVAMP. To assess performance on low-resource African languages across both tasks, we employ AfriMGSM and AfriXNLI. All results are reported as exact-match or accuracy; higher is better.

**MERLIN achieved SOTA on standardized math reasoning benchmarks** Table 1 presents results on MGSM and its algebraic variant MSVAMP. Using the identical Gemma-2 9B backbone, MERLIN achieves 76.2 % average accuracy on MGSM, exceeding MindMerger by +2.8 points and SLAM by +22.8 points.[2] Improvements are consistent across both mid-resource languages (Bn/Th/Sw, +13.7 over Gemma-2) and high-resource languages (+7.7 over MetaMath-Mistral). On MSVAMP, MERLIN retains its lead with 79.2 % average accuracy, outperforming MindMerger by +0.9 points and ex-

[1]The detailed training configuration is stated on Appendix

[2]LangBridge and QAlign scores are omitted in Table 1 because their publicly released implementations do not support the full language set.

| Map | Align | Aug | Spe | eng | fra | amh | ewe | hau | ibo | kin | lin | lug | orm | sna | sot | swa | twi | wol | xho | yor | zul | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ✓ | × | ✓ | × | 58.0 | 50.8 | 45.2 | 32.4 | 44.8 | 34.4 | 43.2 | 38.4 | 36.4 | **40.0** | 44.8 | 36.8 | 53.2 | 24.4 | 21.2 | 33.6 | 34.8 | 40.4 | 39.60 |
| × | ✓ | ✓ | ✓ | **82.8** | **78.0** | 52.8 | 34.0 | 60.4 | 50.4 | 52.0 | 47.2 | **46.8** | 24.0 | 52.8 | 53.2 | 73.6 | 34.8 | 23.6 | 43.2 | 53.6 | 52.4 | 50.87 |
| ✓ | × | ✓ | ✓ | 86.0 | 81.2 | **59.6** | 37.2 | 59.6 | 50.8 | **52.8** | 45.2 | 44.4 | 30.4 | 54.0 | 53.6 | 73.2 | 36.8 | 23.6 | 48.8 | 53.6 | 54.0 | 52.49 |
| ✓ | ✓ | × | ✓ | 77.2 | 64.0 | 26.8 | 16.0 | 30.8 | 20.0 | 28.0 | 19.2 | 20.3 | 13.9 | 24.7 | 24.5 | 33.5 | 16.8 | 10.8 | 22.3 | 24.5 | 24.7 | 27.67 |
| ✓ | ✓ | ✓ | × | 80.0 | 73.2 | 57.2 | 35.6 | 56.8 | 48.0 | 57.2 | 44.4 | 45.2 | 30.4 | 57.6 | 55.2 | 74.0 | 34.0 | 24.0 | 48.4 | 45.2 | 50.4 | 50.93 |
| ✓ | ✓ | ✓ | ✓ | 82.0 | 74.0 | 59.2 | **41.2** | **60.8** | **52.0** | 52.0 | **49.6** | **46.8** | 32.0 | **58.8** | **56.0** | **76.8** | **37.6** | **26.0** | **51.2** | **54.0** | **55.6** | **53.64** |

Table 4: Ablation study: each row removes one stage (×) while keeping the others (✓). Map = General Bilingual Mapping; Align = Question Alignment; Aug = Task-aware Augmentation; Spe = Task Specialization Stage. The first row corresponds to the MindMerger experiment setting.

ceeding the best open baseline by more than 15 points. These gains indicate that the two-stage alignment not only transfers arithmetic skills but also generalises to algebraic word problems.

**MERLIN achieved a wider performance gap in math reasoning for LRLs** Table 2 focuses on sixteen truly low-resource African languages. With the strong Gemma-2 9B+NLLB-600M configuration, MERLIN attains 50.6 % average accuracy, surpassing MindMerger by +12.9 points, SLAM by +11.6 points, and QAlign-LoRA by +16.9 points. Notably, MERLIN outruns the closed GPT-4o-mini by +15.2 points and comes within 2.0 points of GPT-4o despite the latter's much larger parameter budget. Ablation confirms (Table 4) that the task-aware augmentation sub-stage is decisive, yielding gains of more than ten points when added to the connector training.

**Reasoning capabilities generalizes to NLU tasks** Table 3 reports accuracy on sixteen African languages plus English and French on NLI task. MERLIN attains 68.7 % average accuracy on the African subset, improving on MindMerger by +1.6 points and on the base Gemma–2 9B model by +28.4 points, and marginally surpassing the closed GPT-4o (64.3 %). The gains are most pronounced for *Amharic* (+5.6), *Oromo* (+4.6), *Hausa* (+2.5), and *Kinyarwanda* (+2.0), confirming the value of the cross-lingual mapping curriculum in truly low-resource settings. The overall improvement across thirteen of the sixteen African languages indicates that MERLIN successfully redistributes modelling capacity toward languages that benefit the most from additional cross-lingual supervision, while maintaining competitive performance on those already well supported.

## 5 Analysis of Language alignment and Effects on multilingualism

**Effectiveness in Low and Mid-resource reasoning Settings** Across all three evaluation results

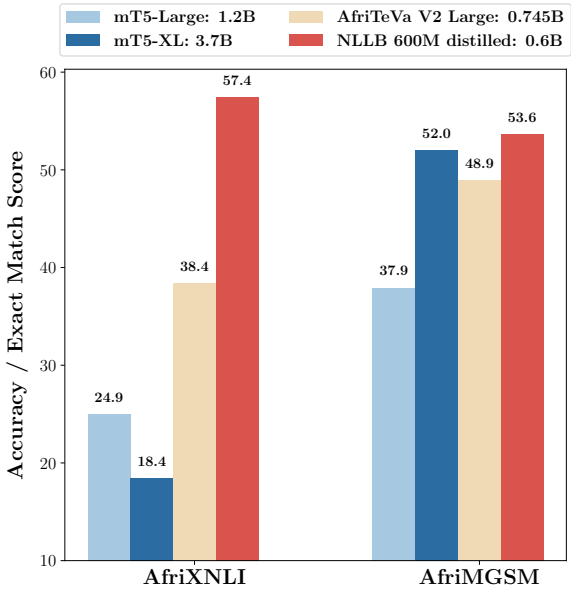


Figure 2: Comparison of MERLIN performance across five different multilingual encoders and Gemma 2 9B LLM.

(Table 1, 2), MERLIN delivers its greatest improvements in the settings with lower resource settings. On MGSM, the model improves over MINDMERGER by **+6.5** accuracy points for the mid-resource group (Bn, Th, Sw), while the corresponding gain for the seven high-resource languages is only **+1.3**. The same pattern holds for MSVAMP (**+2.2** vs. **+0.4**). The advantage widens on the 16-language AFRIMGSM benchmark, where MERLIN averages a **+12.9**-point delta and delivers double-digit improvements in ten languages. We attribute the larger mid and low-resource gains to the combination of *Question Alignment* (Sub-stage Ib) and the PEFT–based *Task Specialization* of Stage II. The former links each non-English question to an English prompt of identical structure, improving cross-lingual grounding, while the latter tunes a small set of low-rank weights to consolidate this grounding inside the decoder without disturbing the frozen backbone. Evidence comes from the ablations in Table 4: disabling Ib (MAP ✓ ALIGN × AUG ✓ SPE ✓) or Stage II (MAP ✓ ALIGN ✓ AUG ✓ SPE ×) re-

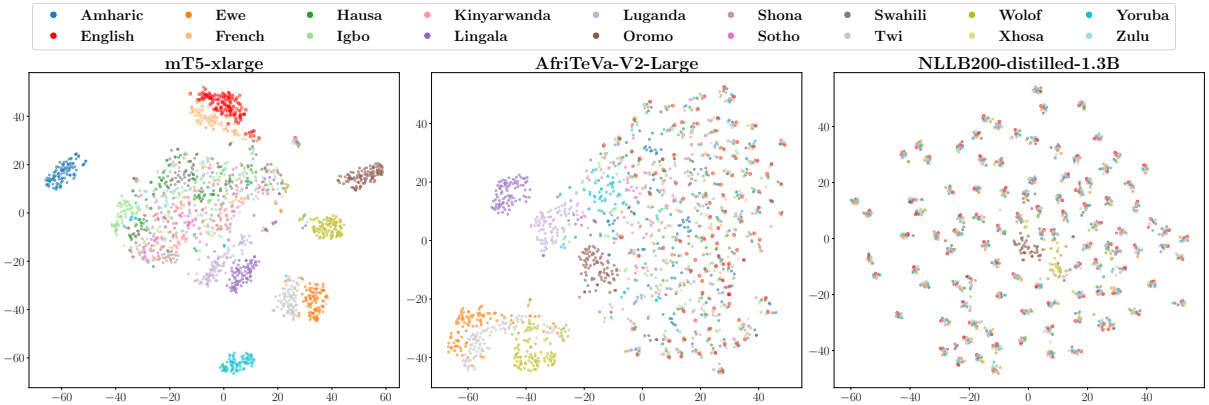


Figure 3: T-SNE visualizations of sentence embeddings for mT5-xl, AfriTeVa-V2-Large, and NLLB200-distilled-1.3B, showing how LRL data impacts cross-lingual alignment. More data reduces clustering, bringing low-resource languages closer to English, enhancing transfer learning.

duces the overall score from 53.6 to 50.9 and 52.5, respectively, with the steepest drops concentrated in the low-resource languages.

**Encoders rich in LRL data sharpen cross-Lingual alignment** A key factor in multilingual model merging is how different encoder choices affect the representation of LRLs in the mapping layer. To isolate this effect, we kept the MERLIN architecture unchanged across all experiments and varied only the multilingual encoder. As shown in Figure 3, when the proportion of LRL text is small, non-English data—especially from LRLs—forms distinct clusters far from English in the LLM's original embedding space. This separation prevents the model from sharing or transferring internal reasoning paths learned predominantly through English data. Conversely, incorporating more low-resource data enables the mapping layer to align different language embeddings, including those from LRLs, more closely with English. By bringing LRLs into the same representational space as English, the LLM can more effectively leverage its inherent reasoning capabilities for LRL tasks, rather than leaving them isolated at the margins of the embedding space. Consequently, increasing low-resource language datasets during encoder training promotes cross-lingual knowledge transfer and enhances representation consistency across languages.

**Middle Decoder Layers Reach Peak Cross-Lingual Alignment** Figure 4 reports sentence-level retrieval@5 for every decoder layer of GEMMA2-9B-IT, MINDMERGER, and MERLIN.

Retrieval@5 is computed by taking each English sentence in the FLORES-200 dev set as a query, ranking all candidate sentences in the paired target

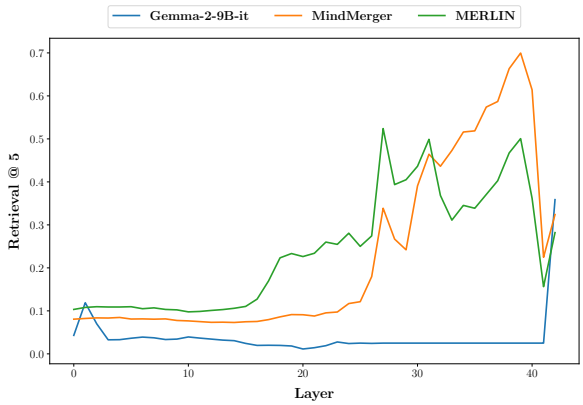


Figure 4: **Layer-wise cross-lingual retrieval@5.** Scores are averaged over English and 16 African languages. MERLIN achieves the highest mid-layer alignment, whereas MindMerger peaks only in the final decoder blocks.

language by the cosine similarity of their layer-wise embeddings, and then checking whether the gold translation appears among the top-5 nearest neighbours. The score—averaged over 200 sentence pairs and 15 African languages—therefore reflects the probability that a representation keeps the correct cross-lingual match within a five-item shortlist, providing a direct proxy for how language-agnostic the layer's embedding space is.

From layers 18–28 MERLIN reaches its peak sentence–level retrieval@5 (~0.30). The rise coincides with the PEFT fine–tuning in Stage II, which adjusts a small set of low-rank weights on top of the frozen LLM while keeping the connector $f_{\sigma^\star}$ fixed. This adaptation strengthens the multilingual abstractions that naturally emerge in the decoder's middle blocks—the region reported to be the most language-agnostic once parallel supervision is available (Liu and Niehues, 2025). As a result, non-English queries are already embedded

in the English reasoning sub-space before the later, generation-oriented layers are reached, providing a solid cross-lingual scaffold for downstream mathematical reasoning tasks such as AFRIMGSM.

## 6 Related Work

**LLMs in LRLs** LRLs often suffer from limited annotated data and high-quality machine translation resources, creating significant challenges for NLP models (Magueresse et al., 2020; Lee et al., 2022; Khan et al., 2023; Deshpande et al., 2024; Qin et al., 2024; Tonja et al., 2024). Moreover, these languages can exhibit complex morphological structures and orthographic variations (Ghosh et al., 2024; Nzeyimana, 2024; Lopo and Tanone, 2024; Issaka et al., 2024; Lusito et al., 2023), posing further challenges. While multilingual models can handle many languages, they frequently underperform in truly low-resource settings and may even trail behind strong monolingual baselines (Yoon et al., 2024; Du et al., 2024; Blevins et al., 2024). Existing efforts to boost LRL performance focus primarily on dataset creation (Adelani et al., 2021; Muhammad et al., 2022; Adelani et al., 2024a; Yong et al., 2024) or adaptation of pretrained models (Alabi et al., 2022; Wu et al., 2024; Csaki et al., 2023; Schmidt et al., 2024), yet these strategies remain constrained by data scarcity and fail to fully address complex linguistic properties (Urbizu et al., 2023; Wang et al., 2023).

**Multilingual Reasoning.** Recent efforts to enhance multilingual *reasoning* in LLMs include carefully devised prompt engineering to facilitate multi-step inference or chain-of-thought reasoning (Huang et al., 2023; Qin et al., 2023), although these techniques often lose effectiveness with smaller open-source models (Touvron et al., 2023). Another approach involves supervised fine-tuning on translated data (Chen et al., 2024; Zhu et al., 2024), which infuses language or task-specific knowledge into the model. Building upon this, Fan et al. (2025) extend the work of Zhu et al. (2024) by optimizing both the training parameters and the training stages focusing on language-specific neurons. Their improvements lead to higher average performance on multilingual mathematical reasoning benchmarks. However, reliance on machine translation can introduce errors (Shi et al., 2022) and may overlook the LLM's built-in multilingual capabilities. To circumvent these issues, MindMerger (Huang et al., 2024) sidesteps autoregressive translation by directly integrating an external multilingual encoder, preserving the LLM's internal parameters.

## 7 Conclusion

This work introduced MERLIN, a lightweight framework for strengthening multilingual mathematical reasoning and natural-language inference in large language models. A two-stage curriculum first learns a compact connector that projects multilingual encoder states into the frozen LLM embedding space, then refines the decoder through DoRA-based task specialization. The approach preserves the original English reasoning capabilities while substantially improving performance in both mid- and low-resource languages.

Our analysis shows that model merging enhances cross-lingual alignment across layers compared to using LLMs alone. In particular, our framework, MERLIN, improves alignment within the model's internal reasoning subspace, yielding stronger performance on reasoning tasks. We also find that the choice of encoder is critical for low-resource languages: encoders with greater exposure to LRL data produce representations that align more closely with the LLM's internal language (often English), thereby supplying richer cross-lingual signals for model merging.

Empirical evaluation on four benchmarks demonstrates that MERLIN outperforms strong open-source systems—including MindMerger, SLAM, and LangBridge—on MGSM, MSVAMP, AfriMGSM and AfriXNLI, and narrows the gap to closed models such as GPT-4o.

## 8 Limitations

MERLIN relies on automatically translated corpora for all three training stages, assuming that these translations form a reliable semantic bridge between the target language and English. In practice, translation quality is uneven. Even with a strong multilingual model such as the distilled NLLB 600M encoder, we observe systematic errors for certain languages—most notably Oromo—where mistranslations propagate through the mapping layer and erode downstream accuracy. Because the present pipeline admits every machine-translated sentence without verification, noisy lexical choices and hallucinated numerals enter the training signal unchecked. Mitigating this weakness will require explicit data filtering.

A second limitation is the narrow scope of our evaluation. The experiments target two problem families—grade-school mathematics and natural-language inference—selected for their diagnostic value in cross-lingual reasoning. Tasks that demand other forms of structured reasoning, such as program synthesis, commonsense QA, or multi-document summarization, remain unexplored.

Finally, our methodology trains a separate MERLIN instance for each benchmark. This design clarifies the contribution of the mapping curriculum but precludes parameter sharing across tasks and prevents the model from exploiting potential synergies between, for example, symbolic arithmetic and commonsense inference. A unified multi-task variant would be more practical in deployment and could yield additional gains through cross-task transfer, although its optimization dynamics are not yet understood.

## References


David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, and 42 others. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024a. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, and et al. 2024b. Irokobench: A new benchmark for african languages in the age of large language models. *Preprint*, arXiv:2406.03368.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, and 1 others. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10822–10837, Miami, Florida, USA. Association for Computational Linguistics.

Happy Buzaaba, Alexander Wettig, David Ifeoluwa Adelani, and Christiane Fellbaum. 2025. Lugha-llama: Adapting large language models for african languages. *arXiv preprint arXiv:2504.06536*.

Nuo Chen, Zinan Zheng, Ning Wu, Ming Gong, Dongmei Zhang, and Jia Li. 2024. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *Preprint*, arXiv:2310.20246.

Zoltan Csaki, Pian Pawakapan, Urmish Thakker, and Qiantong Xu. 2023. Efficiently adapting pretrained language models to new languages. *Preprint*, arXiv:2311.05741.

Tejas Deshpande, Nidhi Kowtal, and Raviraj Joshi. 2024. Chain-of-translation prompting (cotr): A novel prompting technique for low resource languages. *Preprint*, arXiv:2409.04512.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Xinrun Du, Zhouliang Yu, Songyang Gao, Ding Pan, Yuyang Cheng, Ziyang Ma, Ruibin Yuan, Xingwei Qu, Jiaheng Liu, Tianyu Zheng, Xinchen Luo, Guorui Zhou, Wenhu Chen, and Ge Zhang. 2024. Chinese tiny llm: Pretraining a chinese-centric large language model. *Preprint*, arXiv:2404.04167.

Yuchun Fan, Yongyu Mu, Yilin Wang, Lei Huang, Junhao Ruan, Bei Li, Tong Xiao, Shujian Huang, Xiaocheng Feng, and Jingbo Zhu. 2025. Slam: Towards efficient multilingual reasoning via selective language alignment. *Preprint*, arXiv:2501.03681.

Poulami Ghosh, Shikhar Vashishth, Raj Dabre, and Pushpak Bhattacharyya. 2024. A morphology-based investigation of positional encodings. *Preprint*, arXiv:2404.04530.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Preprint*, arXiv:2403.14608.

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

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 *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Zixian Huang, Wenhao Zhu, Gong Cheng, Lei Li, and Fei Yuan. 2024. Mindmerger: Efficient boosting llm reasoning in non-english languages. *Preprint*, arXiv:2405.17386.

Sheriff Issaka, Zhaoyi Zhang, Mihir Heda, Keyi Wang, Yinka Ajibola, Ryan DeMar, and Xuefeng Du. 2024. The ghanaian nlp landscape: A first look. *Preprint*, arXiv:2405.06818.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, and et al. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Muzammil Khan, Kifayat Ullah, Yasser Alharbi, Ali Alferaidi, Talal Saad Alharbi, Kusum Yadav, Naif Alsharabi, and Aakash Ahmad. 2023. Understanding the research challenges in low-resource language and linking bilingual news articles in multilingual news archive. *Applied Sciences*, 13(15).

En-Shiun Annie Lee, Sarubi Thillainathan, Shravan Nayak, Surangika Ranathunga, David Ifeoluwa Adelani, Ruisi Su, and Arya D. McCarthy. 2022. Pre-Trained Multilingual Sequence-to-Sequence Models: A Hope for Low-Resource Language Translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in fine-tuned llms. *Preprint*, arXiv:2502.14830.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *Preprint*, arXiv:2402.09353.

Joanito Agili Lopo and Radius Tanone. 2024. Constructing and expanding low-resource and underrepresented parallel datasets for indonesian local languages. *Preprint*, arXiv:2404.01009.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10748–10772, Miami, Florida, USA. Association for Computational Linguistics.

Stefano Lusito, Edoardo Ferrante, and Jean Maillard. 2023. Text normalization for low-resource languages: the case of ligurian. *Preprint*, arXiv:2206.07861.

Xuetao Ma, Wenbin Jiang, and Hua Huang. 2025. Problem-solving logic guided curriculum in-context learning for LLMs complex reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8394–8412, Vienna, Austria. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alipio Jeorge, and Pavel Brazdil. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *Preprint*, arXiv:2201.08277.

Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, and 1 others. 2025. Sea-lion: Southeast asian languages in one network. *arXiv preprint arXiv:2504.05747*.

Antoine Nzeyimana. 2024. Low-resource neural machine translation with morphological modeling. *Preprint*, arXiv:2404.02392.

Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Esther Adenuga, David Ifeoluwa Adelani, and Jimmy Lin. 2025. Improving multilingual math reasoning for african languages. *arXiv preprint arXiv:2505.19848*.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David Ifeoluwa Adelani. 2024. How good are large language models on african languages? *Preprint*, arXiv:2311.07978.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. 2023. Better quality pre-training data and t5 models for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168, Singapore. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,

Diogo Almeida, Janko Altenschmidt, Sam Altman, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2024. Multilingual large language model: A survey of resources, taxonomy and frontiers. *Preprint*, arXiv:2404.04925.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural Machine Translation for Low-resource Languages: A Survey. *ACM Comput. Surv.*, 55(11):229:1–229:37.

Zhiwen Ruan, Yixia Li, He Zhu, Longyue Wang, Weihua Luo, Kaifu Zhang, Yun Chen, and Guanhua Chen. 2025. Layalign: Enhancing multilingual reasoning in large language models via layer-wise adaptive fusion and alignment strategy. *Preprint*, arXiv:2502.11405.

Fabian David Schmidt, Philipp Borchert, Ivan Vulić, and Goran Glavaš. 2024. Self-distillation for model stacking unlocks cross-lingual NLU in 200+ languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6724–6743, Miami, Florida, USA. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum Learning: A Survey. *Int. J. Comput. Vision*, 130(6):1526–1565.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, and et al. 2024a. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, and et al. 2024b. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, and et al. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Atnafu Tonja, Fazlourrahman Balouchzahi, Sabur Butt, Olga Kolesnikova, Hector Ceballos, Alexander Gelbukh, and Thamar Solorio. 2024. NLP progress in indigenous Latin American languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6972–6987, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Kosei Uemura, Mahe Chen, Alex Pejovic, Chika Maduabuchi, Yifei Sun, and En-Shiun Annie Lee. 2024. AfriInstruct: Instruction tuning of African languages for diverse tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13571–13585, Miami, Florida, USA. Association for Computational Linguistics.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, Rodrigo Agerri, and Aitor Soroa. 2023. Scaling laws for BERT in low-resource settings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7771–7789, Toronto, Canada. Association for Computational Linguistics.

Yudong Wang, Chang Ma, Qingxiu Dong, Lingpeng Kong, and Jingjing Xu. 2023. A challenging benchmark for low-resource learning. *Preprint*, arXiv:2303.03840.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *Preprint*, arXiv:1704.05426.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation. *Preprint*, arXiv:2401.06468.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2024. Lexc-gen: Generating data for extremely low-resource languages with large language models and bilingual lexicons. *Preprint*, arXiv:2402.14086.

Dongkeun Yoon, Joel Jang, Sungdong Kim, Seungone Kim, Sheikh Shafayat, and Minjoon Seo. 2024. LangBridge: Multilingual reasoning without multilingual supervision. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7502–7522, Bangkok, Thailand. Association for Computational Linguistics.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2023. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024. Metamath: Bootstrap your own mathematical questions for large language models. *Preprint*, arXiv:2309.12284.

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

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, and et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

## A Dataset-Size Ablation

With the full sweep of dataset sizes in Table 5, a clear elbow appears at **1 000** translated sentence pairs per language and per stage. At that level MERLIN attains its highest averages—78.3 % on MGSM and 80.4 % on MSVAMP—while using only one-sixth of the data required by the largest setting. A modest increase to 1 500 pairs produces small, mixed fluctuations, and every larger tier either plateaus or declines, a pattern we attribute to domain mismatch and over-fitting of the lightweight connector once the easiest bilingual correspondences have been captured. Even the 500-pair condition is competitive (75.2 % / 80.1 %), suggesting that the first few hundred high-quality translations already establish a usable cross-lingual bridge. In terms of compute, the 1 000-pair configuration is exceptionally light: all three connector stages plus DoRA adaptation complete in just under one hour on a single A100-80 GB GPU (batch size = 1), confirming that MERLIN can be trained end-to-end in resource-constrained settings without sacrificing downstream accuracy.

## B Complete Encoder Ablation Study

Table 6 compares four multilingual encoders (MT5-LARGE, MT5-XL, AFRITEVA-V2-LARGE, and NLLB-600 M DISTILLED) while keeping the MERLIN decoder, connector, and all training settings fixed. Two evaluation suites are considered: AFRIXNLI (classification) and AFRIMGSM (mathematical reasoning).

On AFRIXNLI, the distilled NLLB encoder delivers the highest average accuracy (55.6 %), improving over the best mT5 variant by more than 30 points and over AFRITEVA-V2 by 20 points. Gains are especially large for the most data-scarce languages such as Amharic (+49.5 vs. mT5-Large) and Oromo (+27.5 vs. AfriTeVa-v2), confirming that broad pre-training coverage is critical for cross-lingual alignment.

For AFRIMGSM, NLLB again attains the highest average (51.5 %), edging out MT5-XL by two points. While mT5-XL excels on high-resource pivots (English, French) and on Swahili, it underperforms on several low-resource languages (e.g. Ewe, Wolof). AFRITEVA-V2 provides competitive results on Swahili and Sotho but lags elsewhere, reflecting its narrower linguistic coverage.

Overall, the results highlight that encoders trained on a broad, balanced corpus—such as NLLB—yield the most reliable cross-lingual representations when coupled with MERLIN, particularly in low-resource African settings. Models with less coverage (MT5-XL) or regional focus (AFRITEVA-V2) may excel on individual languages but fall short in overall robustness.

## C Ablation of Training Mapping Layers

Table 7 varies the expressiveness of the mapping head while holding every other component of MERLIN fixed. A linear projection or a single hidden layer yields only modest averages and large per-language variance. Adding a second hidden layer (9.45 M parameters) stabilises performance, but the decisive improvement comes from inserting a residual skip: the resulting two-layer *Residual MLP* attains 77.6 % on MGSM and 81.1 % on MSVAMP, the best figures in both suites. Expanding to three hidden layers (13.6 M parameters) reverses the gain, dropping accuracy by 1–2 points and confirming that extra depth overfits the limited translation supervision.

The residual two-layer MLP therefore offers the strongest accuracy–capacity trade-off, adding fewer than ten million parameters while outperforming shallower and deeper variants across mid- and high-resource languages. At the same time the gap between 77.6 % and the theoretical ceiling on MGSM suggests room for architectural exploration—e.g. gated skips, attention-based adapters, or language-conditioned hypernetworks—that could push the mapping layer beyond its current residual design.

## D Ablation of Training Adapters

Table 8 contrasts two parameter-efficient adapters applied to the decoder. LoRA (Hu et al., 2022)

| Model / Size | Bn | Th | Sw | Ja | Zh | De | Fr | Ru | Es | En | Mrl. | Hrl. | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MGSM** | | | | | | | | | | | | | |
| 500 | 72.8 | 73.6 | 71.6 | 72.4 | 73.2 | 74.8 | 75.6 | 79.6 | 76.8 | 82.0 | 72.7 | 76.3 | 75.2 |
| 1000 | **76.4** | 77.6 | 74.4 | 71.6 | **76.0** | **78.4** | 77.6 | **81.6** | 83.6 | **85.6** | 76.1 | **79.2** | **78.3** |
| 1500 | 75.6 | 74.4 | 71.6 | 72.8 | 76.4 | 74.4 | 75.6 | 78.8 | **85.2** | 84.8 | 73.9 | 78.3 | 77.0 |
| 3000 | 74.8 | 76.0 | 77.6 | 71.6 | 72.8 | 73.2 | 74.8 | 78.4 | 77.6 | **85.6** | 76.1 | 76.3 | 76.2 |
| 4500 | 74.8 | 77.6 | 72.8 | **74.0** | 74.4 | 75.6 | 75.6 | 78.4 | 82.8 | 84.0 | 75.1 | 77.8 | 77.0 |
| 9000 | 74.8 | 78.4 | 74.8 | **74.0** | 72.4 | 76.0 | 74.8 | 77.6 | 81.2 | 83.6 | 76.0 | 77.1 | 76.8 |
| 18000 | 76.0 | **79.6** | **79.2** | 73.2 | 72.8 | 76.8 | **79.2** | 78.0 | 82.8 | 83.6 | **78.3** | 78.1 | 78.1 |
| **MSVAMP** | | | | | | | | | | | | | |
| 500 | 72.6 | 75.3 | 77.8 | 80.7 | 81.2 | 82.8 | 82.5 | 81.0 | 83.1 | 83.5 | 75.2 | 82.1 | 80.1 |
| 1000 | **73.0** | 75.7 | **79.4** | 79.7 | **82.2** | **82.9** | **82.7** | **81.6** | **82.8** | 84.4 | **76.0** | **82.3** | **80.4** |
| 1500 | 71.8 | **76.5** | 78.4 | **81.1** | 80.7 | 82.3 | 82.0 | 81.4 | 83.3 | **85.2** | 75.6 | 82.3 | 80.3 |
| 3000 | 72.1 | 76.1 | 77.4 | 78.7 | 79.8 | 81.5 | 81.8 | 79.8 | 81.8 | 83.3 | 75.2 | 81.0 | 79.2 |
| 4500 | 70.1 | 75.6 | 78.9 | 78.7 | 79.1 | 80.7 | 80.9 | 79.8 | 81.5 | 83.3 | 74.9 | 80.6 | 78.9 |
| 9000 | 71.6 | 76.0 | 77.1 | 78.7 | 79.5 | 81.3 | 81.3 | 80.4 | 81.6 | 81.9 | 74.9 | 80.7 | 78.9 |
| 18000 | 71.4 | 74.7 | 74.6 | 77.1 | 77.2 | 82.4 | 80.3 | 79.1 | 80.9 | 85.3 | 73.6 | 80.3 | 78.3 |

Table 5: MERLIN performance on MGSM and MSVAMP across synthetic-training set sizes. Bn/Th/Sw are mid-resource (MRL); the remaining seven languages are high-resource (HRL).

| Model | eng | fra | amh | ewe | hau | ibo | kin | lin | lug | orm | sna | sot | swa | twi | wol | xho | yor | zul | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **AfriXNLI** | | | | | | | | | | | | | | | | | | | |
| mT5-Large | 22.8 | 37.8 | 9.7 | 40.0 | 33.5 | 14.5 | 29.0 | 22.0 | 13.5 | 25.7 | 35.3 | 39.0 | 38.2 | 9.5 | 25.7 | 16.8 | 19.0 | 17.0 | 24.3 |
| mT5-XL | 24.8 | 24.8 | 18.3 | 25.0 | 21.3 | 7.0 | 25.3 | 26.8 | 13.0 | 9.0 | 21.0 | 19.2 | 16.5 | 12.7 | 22.2 | 14.3 | 9.8 | 20.2 | 17.6 |
| AfriTeVa V2 Large | 61.0 | **66.0** | 14.3 | 45.2 | 42.3 | 31.5 | 40.3 | 29.3 | 31.2 | 30.7 | 44.5 | 42.8 | 40.7 | 26.2 | 34.0 | 39.5 | 33.0 | 38.3 | 35.2 |
| NLLB 600M distilled | **83.2** | 62.0 | **69.2** | **47.9** | **69.0** | **43.9** | **68.4** | **32.2** | **47.2** | **53.2** | **63.4** | **66.7** | **60.5** | **45.0** | **52.5** | **56.7** | **54.5** | **58.9** | **55.6** |
| **AfriMGSM** | | | | | | | | | | | | | | | | | | | |
| mT5-Large | 80.0 | 68.0 | 49.6 | 13.6 | 54.0 | 27.6 | 40.0 | 23.2 | 31.2 | 18.8 | 40.0 | 49.2 | 64.4 | 12.4 | 7.6 | 39.6 | 27.6 | 35.6 | 34.6 |
| mT5-XL | **82.8** | **80.0** | **62.0** | 28.4 | **64.0** | 50.4 | **52.4** | 41.2 | 41.2 | 42.8 | **62.0** | **59.2** | **79.2** | 24.0 | 15.2 | 49.2 | 47.2 | 55.2 | 49.5 |
| AfriTeVa V2 Large | 85.2 | 77.2 | 60.0 | 11.6 | 62.0 | 47.6 | **52.4** | 31.6 | 40.8 | **51.2** | 53.6 | 55.2 | 77.2 | 11.6 | 8.8 | 46.8 | 53.6 | 54.0 | 46.0 |
| NLLB 600M distilled | 82.0 | 74.0 | 59.2 | **41.2** | 60.8 | **52.0** | 52.0 | **49.6** | **46.8** | 32.0 | 58.8 | 56.0 | 76.8 | **37.6** | **26.0** | **51.2** | **54.0** | **55.6** | **51.5** |

Table 6: Accuracy of four multilingual models on **AfriXNLI** and **AfriMGSM**. Avg is computed over the 16 African languages (all columns except *eng* and *fra*).

| Mapping head | # Parm | Bn | Th | Sw | Ja | Zh | De | Fr | Ru | Es | En | Mrl. | Hrl. | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MGSM** | | | | | | | | | | | | | | |
| Linear | 3.68 M | 72.4 | 75.2 | 70.4 | **75.2** | **74.0** | **79.2** | 77.2 | **81.2** | 80.8 | 85.2 | 72.7 | **78.9** | 77.08 |
| 1-Layer MLP | 3.68 M | 65.6 | 68.8 | 71.2 | 68.8 | 62.8 | 71.6 | 73.6 | 78.0 | 77.2 | 82.0 | 68.5 | 73.4 | 71.5 |
| 2-Layer MLP | 9.45 M | 74.8 | 76.0 | **77.6** | 71.6 | 72.8 | 73.2 | 74.8 | 78.4 | 77.6 | **85.6** | 76.1 | 76.3 | 76.2 |
| 3-Layer MLP | 13.64 M | 74.8 | 76.4 | 69.6 | 70.0 | 71.2 | 74.0 | 72.0 | 76.8 | 79.2 | 79.6 | 73.6 | 74.7 | 74.4 |
| Residual MLP | 9.45 M | **75.2** | **80.0** | 77.2 | 73.6 | 70.8 | 78.8 | **78.0** | 79.2 | **82.4** | 80.8 | **77.5** | 77.7 | **77.6** |
| **MSVAMP** | | | | | | | | | | | | | | |
| Linear | 3.68 M | 72.8 | 75.2 | 77.8 | 79.5 | 79.9 | 83.7 | 81.4 | 80.3 | 82.0 | 83.0 | 75.3 | 81.4 | 79.6 |
| 1-Layer MLP | 3.68 M | 69.4 | 75.8 | 77.3 | 79.3 | 79.9 | 81.1 | 81.3 | 80.0 | 81.7 | 82.4 | 74.2 | 80.8 | 78.8 |
| 2-Layer MLP | 9.45 M | 72.1 | 76.1 | 77.4 | 78.7 | 79.8 | 81.5 | 81.8 | 79.8 | 81.8 | 83.3 | 75.2 | 81.0 | 79.2 |
| 3-Layer MLP | 13.64 M | 71.0 | 74.0 | 76.7 | 77.3 | 76.1 | 80.8 | 82.0 | 78.7 | 80.6 | 81.5 | 73.9 | 79.6 | 77.9 |
| Residual MLP | 9.45 M | **72.7** | **77.6** | **80.7** | **79.6** | **81.0** | **84.8** | **83.3** | **83.2** | **83.7** | **84.5** | **77.0** | **82.9** | **81.1** |

Table 7: Effect of mapping-head capacity on MGSM and MSVAMP. **Mrl.** = mean over mid-resource languages (Bn, Th, Sw); **Hrl.** = mean over the high-resource set; "–" indicates the score is not available.

injects trainable low-rank update matrices into the linear weights, while all original parameters stay frozen. DoRA (Liu et al., 2024) keeps the same low-rank structure but factorises each weight into a direction and a magnitude, yielding the same footprint yet a more explicit decomposition.

With rank and parameter count held constant, the two adapters behave nearly identically: average accuracy on MGSM differs by only 0.2 pp, and the MSVAMP results coincide after rounding. Performance across mid- and high-resource language subsets mirrors this pattern, indicating that adapter choice has minimal impact under our training recipe.

| Adaptation | # Parm | Bn | Th | Sw | Ja | Zh | De | Fr | Ru | Es | En | Mrl. | Hrl. | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **MGSM** | | | | | | | | | | | | | | |
| LoRA | 18.4 M | 75.2 | 76.4 | 77.2 | 71.6 | 72.4 | 72.4 | 75.6 | 79.2 | 78.0 | 85.6 | 76.3 | 76.4 | 76.4 |
| DoRA | 18.4 M | 74.8 | 76.0 | 77.6 | 71.6 | 72.8 | 73.2 | 74.8 | 78.4 | 77.6 | 85.6 | 76.1 | 76.3 | 76.2 |
| **MSVAMP** | | | | | | | | | | | | | | |
| LoRA | 18.4 M | 71.6 | 76.2 | 77.3 | 78.0 | 79.1 | 81.6 | 82.1 | 80.1 | 82.2 | 83.6 | 75.0 | 81.0 | 79.2 |
| DoRA | 18.4 M | 72.1 | 76.1 | 77.4 | 78.7 | 79.8 | 81.5 | 81.8 | 79.8 | 81.8 | 83.3 | 75.2 | 81.0 | 79.2 |

Table 8: Ablation of LoRA vs. DoRA adapters for MERLIN (Gemma-2 9 B + NLLB-600 M). **Mrl.** = mean over mid-resource languages (Bn, Th, Sw); **Hrl.** = mean over seven high-resource languages; averages are arithmetic means of the shown scores.

## E AmericasNLI Evaluation

To assess whether the mapping curriculum generalises beyond the African and European test beds, we evaluate MERLIN on AMERICASNLI, a three-way NLI benchmark that targets Aymara, Guarani, and Quechua. For MERLIN and the MindMerger baseline, we follow the same data pipeline as for AFRIXNLI: 3,000 English premise–hypothesis pairs are sampled from MULTINLI, translated into each target language with NLLB 600m distilled, and used as supervision for Stage I.b, Stage I.c, and Stage II.[3] The non-adapted LLaMAX2-7B-XNLI (Lu et al., 2024) serves as the baseline decoder-only model. Results are summarised in Table 9.

| Model | Aymara | Guarani | Quechua | Avg. |
|---|---|---|---|---|
| LLaMAX2 XNLI (base) | 41.6 | 43.6 | 40.7 | 42.0 |
| MindMerger | **48.8** | 48.7 | 48.3 | 48.6 |
| MERLIN | 48.4 | **49.3** | **50.5** | **49.4** |

Table 9: Accuracy (%) on AMERICASNLI. Average is computed over the three indigenous languages.

MERLIN outperforms the base model by $+7.4$ accuracy points on average and still outperform past MindMerger by $+0.8$ points. The consistent improvement across all three languages suggests that the cross-lingual alignment learned from African and Indo-European data transfers to typologically distant language families in the Americas.

## F Training Configurations

All parameter-efficient runs (MERLIN, QAlign-LoRA) employ rank-16 LoRA/DoRA adapters injected into the self-attention `q_proj` and `v_proj` projections:

| Hyper-parameter | Value |
|---|---|
| Rank $r$ | 16 |
| LoRA / DoRA $\alpha$ | 32 |
| Dropout | 0.05 |
| Bias | none |
| Target modules | `q_proj`, `v_proj` |

QAlign and SLAM adopt identical settings from the released code [4]. For LangBridge we follow the original implementation, using a single linear projection layer and sampling 100k examples from METAMATHQA training set

Both MERLIN and MindMerger are trained on identical corpora. Stage I.a uses 9,000 generic English–target sentence pairs per language drawn from the NLLB-200 corpus. Stage I.b is supplied with up to 3,000 automatically translated questions per language, also generated with NLLB. For Stage I.c and the subsequent Task-Specialisation stage, task-specific data are employed: the MGSM track relies on the translated METAMATHQA split, whereas the AfriXNLI track is trained with translated MULTINLI.

## G Prompt Templates

LangBridge and SLAM are executed through `lm-eval-harness`; their default prompts are therefore retained and not repeated here. Table 10 shows the exact text used by MERLIN, MindMerger, and QAlign for the two evaluation tracks.

[3] No task-specific data from the original AMERICASNLI development set is seen during training.

[4] https://github.com/NJUNLP/QAlign,https://github.com/fmm170/SLAM

| **Mathematical Reasoning (MGSM, MSVAMP, AfriMGSM)** | |
|---|---|
| **Gemma-2 backbone** | <bos><start_of_turn>user<br>*Below is an instruction that describes a task.*<br>*Write a response that appropriately completes the request.*<br>{query}<br>*Let's think step by step.*<br><end_of_turn><start_of_turn>model |
| **MetaMath-style backbones** | *Below is an instruction that describes a task.*<br>*Write a response that appropriately completes the request.*<br>### Instruction: {query}<br>### Response: Let's think step by step. |
| **Natural-Language Inference (AfriXNLI)** | |
| **All backbones** | **Premise:** {sentence1}<br>**Hypothesis:** {sentence2}<br>**Label:** |

Table 10: Prompt templates used for MERLIN, MindMerger and QAlign. Curly braces denote insertion points for the instance text.