Data-Centric Continual Pre-training for 500+ Languages: A New Bilingual Translation Corpus and Multilingual Models

Shaoxiong Ji[1,2*] **Zihao Li**[3] **Jaakko Paavola**[3] **Hengyu Luo**[3] **Jörg Tiedemann**[1,3]
[1]ELLIS Institute Finland [2]University of Turku [3]University of Helsinki
shaoxiong.ji@utu.fi
{zihao.li, jaakko.paavola, hengyu.luo, jorg.tiedemann}@helsinki.fi

Abstract

This paper investigates a critical design decision in the practice of massively multilingual continual pre-training — the inclusion of parallel data. Specifically, we study the impact of bilingual translation data for massively multilingual language adaptation of the Llama3 family of models to 500 languages. To this end, we construct a bilingual translation corpus named MaLA, containing data from more than 2,500 language pairs. Subsequently, we develop the EMMA-500 Llama 3 suite of four massively multilingual models — continually pre-trained from the Llama 3 family of base models extensively on diverse data mixes up to 671B tokens — and explore the effect of continual pre-training with or without bilingual translation data. Comprehensive evaluation across 7 tasks and 12 benchmarks demonstrates that bilingual data tends to enhance language transfer and performance, particularly for low-resource languages. We open-source the MaLA corpus, EMMA-500 Llama 3 suite artefacts, code, and model generations.

Website: mala-lm.github.io/emma-500-gen2
Models: MaLA-LM/emma-500
Data: MaLA-LM/mala-bilingual-translation-corpus
Evaluation: github.com/MaLA-LM/emma-500

1 Introduction

Large language models (LLMs) pre-trained on massive data have promoted multilingual natural language processing (NLP). However, multilingual models such as BLOOM (Scao et al., 2022) and Llama (Touvron et al., 2023a,b) often struggle with low-resource languages and are still limited in their language coverage (Huang et al., 2023; Sindhujan et al., 2025; Huang et al., 2025). Recent works extend pre-trained LLMs into multiple languages via continual pre-training (CPT). For example, LLaMAX (Lu et al., 2024) and xLLMs-100 (Lai et al., 2024) adopt CPT and instruction fine-tuning to extend existing LLMs into 100 languages, and MaLA-500 (Lin et al., 2024) and EMMA-500 (Ji et al., 2024a) perform continual pre-training (low-rank and full-parameter CPT using Llama 2) to adapt LLMs into 500 languages. Despite these efforts, challenges still remain in adapting LLMs to low-resource languages, especially in a massively multilingual scenario with more than 500 languages.

This paper studies CPT in a massively multilingual setting. Prior work like LlaMAX (Lu et al., 2024) uses both monolingual and parallel texts for CPT in 100 languages, and EMMA-500 Llama 2 (Ji et al., 2024a) uses only monolingual texts in 500 languages. Our primary novelty lies in the massive scaling of CPT to over 500 languages using a specifically compiled bilingual translation corpus (i.e, MaLA), and the study of the comparative effects of continual pre-training with monolingual and bilingual translation data.[1]

Contributions Our contributions are three-fold:

- **DATA:** We compile a bilingual translation corpus for Massive Language Adaptation in more than 2,500 language pairs and 500 languages, namely the **MaLA** translation corpus.
- **MODELS:** We train and release 4 models, namely EMMA-500 Llama 3/3.1 Mono/Bi[2] by continually pre-training of Llama 3 & 3.1 (8B) using both monolingual and bilingual MaLA corpus augmented with diverse data types, up to 671B tokens.
- **EVALUATION:** We conduct a comprehensive evaluation across 7 tasks and 12 benchmarks.

[*] Corresponding author. Work done while at the University of Helsinki.

[1]Monolingual data consists of texts written in a single language. Bilingual translation data, also called *parallel corpora*, comprises pairs of sentences in two different languages that express the same meaning. In this paper, we treat the terms bilingual translation corpora/texts/data, bilingual corpora/texts/data, parallel corpora/texts/data, and bitexts as equivalent.

[2]"Mono" and "Bi" indicate CPT on monolingual (fig. 1b) and bilingual (fig. 1a) mixes, respectively.

Findings of the Association for Computational Linguistics: ACL 2026, pages 18776–18807
July 2-7, 2026

Our empirical evaluation ablates the impact of two diverse data mixes and analyzes gains in task generalization and multilingual robustness.

Evaluation Results and Findings

- CPT using a data mix with bilingual translation data generally exhibits better multilingual performance than a monolingual mix[3], particularly in low-resource languages and in machine translation tasks that directly benefit from parallel data.
- Heavily pre-trained models (e.g., Llama 3 and 3.1) that consume more training tokens than English-centric models (e.g., Llama 2) are more resistant to further adaptation when scaling to include many additional languages.
- As for overall performance, comparing to strong baselines, our `EMMA-500` models are the best at machine translation (Flores200) and competitive at text classification (Taxi1500 and SIB-200) and commonsense reasoning (XCOPA and XStoryCloze).
- `EMMA-500` CPT models exhibit a lower average accuracy on the BELEBLE comprehension benchmark, but they outperform baselines across a greater number of languages.

While multilingual models can achieve broad coverage, perfect uniformity across all tasks and languages remains an unattainable goal. However, we show that multilingual performance and language equality can be pushed forward with parallel training data.

Outline Section 2 presents the data and model training with a newly compiled bilingual translation corpus introduced in Section 2.1, data mixing introduced in Section 2.2, and settings for model training introduced in Section 2.3. Appendices A and B describe the details about the `MaLA` translation corpus and data mixes. We evaluate the resulting models and discuss the evaluation results in Section 3. Detailed evaluation setup and per-benchmark and per-language results are presented in Appendices C and D respectively. We conclude the paper in Section 4. Appendices E and F introduce related work and ethics consideration.

[3]A monolingual mix (fig. 1b) contains monolingual data in different languages but not in the aligned format as parallel data.

2 Data and Model Training

2.1 `MaLA` Translation Corpus

Bilingual translation corpora are language datasets that contain text data in one language aligned with text data in another language. We extend the MaLA (Massive Language Adaptation) corpus (Ji et al., 2024a) by incorporating parallel data in more than 500 languages and 2,500 language pairs. The resulting parallel dataset is named the `MaLA` translation corpus (**`MaLA`** for short), which is suitable for adapting language models in massively multilingual scenarios. The section describes the process of building the `MaLA` translation corpus in a way similar to the MaLA corpus with monolingual texts, but focuses on bilingual texts.

We follow a similar data integration pipeline to the process of compiling the MaLA corpus (Ji et al., 2024a), including extraction, harmonization, language code normalization, and writing system recognition. The bilingual data comes from various sources, including OPUS (Tiedemann, 2012), NLLB (NLLB Team et al., 2022), and Tatoeba (Tiedemann, 2020). The datasets from OPUS are made by an existing compilation: Lego-MT (Yuan et al., 2023). Table 6 in Appendix A.1 shows the data sources for bilingual texts. The main difference in the script recognition, as well as language code conversion, with the bilingual corpora, is in the form of the label; we obtain a label in the form of a language pair, e.g., eng_Latn-zho_Hani.

Language Code Normalization Language code normalization converts various language codes into a standardized format to ensure consistency and compatibility across different systems and applications. With the bilingual corpora, we face similar issues as with the monolingual ones when converting language denotations given in OPUS[4] to ISO 639-3 language codes. Moreover, with bilingual corpora, we want to specify dialects based on the ISO 3166-1 alpha-3 standard. The recognition and handling of language codes are based on the following procedure:

- If the language code of a dataset provided by OPUS matches a language code in ISO-639-3, then we consider it such.
- If the language code does not match one in ISO-639-3, we use the `langcodes` package[5] to convert it to ISO-639-3.

[4]https://github.com/Helsinki-NLP/OPUS
[5]https://pypi.org/project/langcodes/

- If the above steps fail, we assign "unknown" as the language code.

Writing System Recognition We implement writing system identification following ISO 15924 standards using the GlotScript library (Kargaran et al., 2023). For script detection, we analyze 100-line samples by default, reverting to first-line analysis when standard detection fails—a known limitation affecting both our bilingual and monolingual corpora (Ji et al., 2024a). We do not classify a dataset into multiple scripts, even in cases of code-mixing, where multiple scripts are used.

Data Cleaning The bilingual corpus compilation faces significant quality variability, such as noisy sentence pairs. The Lego-MT dataset (Yuan et al., 2023) applies some data cleaning, including deduplication, removing missing translations, and length mismatching. For other data sources, we add various procedures, including dataset-specific cleaning and deduplication, after integrating all data sets into our collection. We eliminate lines that contain the exact word or character repeated more than five consecutive times. This problem appears in particular in the Tatoeba parallel training data (Tiedemann, 2020) and in the majority of these cases is erroneous. We use OpusFilter (Aulamo et al., 2020) for deduplication of data points.

Key Statistics After pre-processing and cleaning, we obtain the MaLA translation corpus in 2,507 language pairs. Table 1 shows the total number of whitespace-separated tokens and the number of language pairs across different resource categories. Compared with Lego-MT and NLLB, MaLA has a similar number of language pairs but more tokens. We categorize language pairs into 5 resource levels based on token counts: high-resource (>1B), medium-high (>500M), medium (>100M), medium-low (>10M), and low (>1M). Different from the monolingual MaLA, we add two categories of "very high" and "very low" resources in the resource level classification, i.e., very high-resource pairs (>10B) and very low-resource (<1M). In total, there are more than 426B tokens in the MaLA translation corpus.[6] We further sample the MaLA translation corpus for continued training LLMs, considering a balanced corpus size and language coverage.

Categories	Lego-MT		NLLB		MaLA	
	Pairs	Tokens	Pairs	Tokens	Pairs	Tokens
very high	4	5.1E+10	4	2.8E+10	4	8.5E+10
high	51	1.4E+11	51	1.2E+11	83	2.1E+11
medium-high	22	1.5E+10	65	4.5E+10	67	4.7E+10
medium	75	1.8E+10	264	6.1E+10	281	6.4E+10
medium-low	113	4.1E+09	480	2.0E+10	508	2.0E+10
low	350	1.4E+09	491	1.7E+09	655	2.5E+09
very low	1893	1.7E+08	1154	1.3E+08	909	1.8E+08
sum	2508	2.2E+11	2509	2.8E+11	2507	4.3E+11

Table 1: Key statistics of the MaLA translation corpus and comparison with existing parallel corpora.

2.2 Data Mixing

We blend the data compiled in MaLA with multilingual non-parallel data obtained from the cleaned and deduplicated MaLA corpus (Ji et al., 2024a) along with texts selectively sourced from factual and high-quality domains with a view to retaining the knowledge acquired during the *annealing* phase of pre-training (Hu et al., 2024). We mirror existing practice and sample books and scientific papers (Soldaini et al., 2024) along with instruction-like (Maini et al., 2024) data. Following Ji et al. (2024a), we use scientific papers and books from CSL (Li et al., 2022), pes2o (Soldaini and Lo, 2023), and free e-books from the Gutenberg project[7] compiled by Faysse (2023). Multilingual instruction data is sourced from the training set of xp3x (Crosslingual Public Pool of Prompts eXtended)[8] and the Aya collection[9]. Finally, we augment our mixes with code and code-adjacent procedural text due to its demonstrable benefits towards reasoning and entity-tracking (Ruis et al., 2025; Petty et al., 2024) (details in Appendix B.1).

We manually mix up different types of data to balance the language coverage across different levels of resources and types while ensuring that low- and medium-resource languages remain overrepresented. Figure 1 shows the composition categorized by language resources according to the number of tokens. Notably, medium-resource language pairs contribute the majority of bilingual data, and medium-high-resource languages are the largest category for monolingual and instruction data.

[6]Note that whitespace-based token counting is not accurate for a language where words are not typically separated by spaces. In these languages, the absence of whitespace makes it challenging to determine token boundaries, leading to inaccurate token counts when using whitespace as a delimiter. We use whitespace as the delimiter because of its efficiency in processing text.

[7]https://www.gutenberg.org/

[8]https://hf.co/datasets/CohereForAI/xP3x

[9]https://hf.co/datasets/CohereForAI/aya_collection_language_split

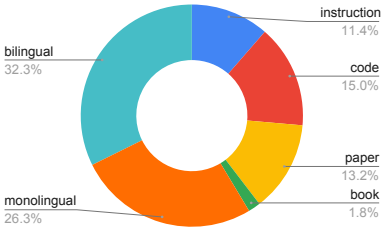
Two Data Mixes We make two data mixes to ablate the effect of incorporating bilingual texts into continual pre-training. The first data mix is a **bilingual mix** (Figure 1b), which incorporates various data types. The second is a **monolingual mix** (Figure 1b), which is derived from the bilingual mix but specifically omits any bilingual data, focusing solely on a subset with monolingual texts per document. Detailed data statistics of these two mixes are presented in Table 8.

Most translation data are aligned sentences, and there is not sufficient document-level aligned data publicly available to be collected in the MaLA corpus, especially in the massively multilingual scenario. When using translation data, we concatenate the source and target language texts in a specific format to form a chunk of pairs in the same language pair. For every ten samples, i.e., sentence pairs, to make a document for training,[10] the format is structured as follows:

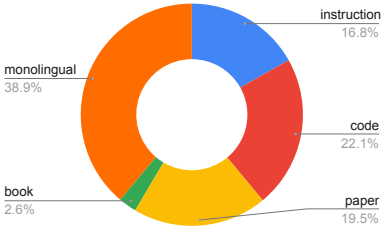
```
[{src_lang_code}]: {src_text} [{
    tgt_lang_code}]: {tgt_text} \n
# 8 lines for 8 samples
[{src_lang_code}]: {src_text} [{
    tgt_lang_code}]: {tgt_text}
```

In this format, the notation $\{\cdot\}$ denotes the variables for source and target language codes and texts. This method allows us to work with pseudo-document-level data and clearly delineate between the source and target languages, facilitating better processing and understanding of the translation data without switching between multiple languages within the context window. By organizing the data this way, we facilitate the model to learn from the relationships between the two languages effectively. This structured approach ensures clarity and consistency in how bilingual data is presented, making it easier to process and analyze. We do not split parallel sentences into independent monolingual strings. Prior literature (Lu et al., 2024) suggests that this degrades translation performance; thus, our method prioritizes optimal practical configurations for multilingual adaptation.

Figure 1 shows the composition of our two data mixes. In the monolingual mix, monolingual web-crawled text is the largest data type, as its name suggests. The bilingual mix incorporates a considerably bigger portion of bilingual texts, 6% more than the monolingual texts. Continual training on these two mixes facilitates the adaptation of LLMs to massively multilingual languages and analyzes the effect of scaling massively multilingual training using bilingual data.



(a) Data mix 1: bilingual



(b) Data mix 2: monolingual

Figure 1: Two date mixes and their composition. The bilingual mix includes all types of data. The monolingual mix consists of a subset of the bilingual mix that excludes bilingual data.

Data Selection & Heuristics Our data mixes in Figure 1 are constructed using a combination of domain expertise and empirical heuristics. We prioritized typological diversity and language coverage, specifically targeting the "long tail" of underrepresented languages. High-quality domains (e.g., books, academic papers, and code) are prioritized to retain the knowledge and enhance reasoning capabilities.

We draw on our knowledge and understanding of the domain to create effective data combinations to balance the language resources and represent different text types. This allows for a reasonable selection that can capture the complexities of the data landscape and is aligned with the goals for massively multilingual adaptation. From an algorithmic perspective, grid search or other similar methods could be used to explore the space of possible data distributions. However, algorithmic search involves systematically exploring a predefined set of hyperparameter values by evaluating all possible combinations through model training on the searched mixes, which can become computationally expensive and time-consuming, especially as the number of parameters increases. This makes

[10]The choice of 10 is inspired by "yí mù shí háng", a Chinese idiom that literally translates to "one glance ten lines", means that someone reads very quickly and efficiently.

it challenging to apply algorithms like grid search effectively in scenarios where the data mix needs to be optimized, which would require thousands of training runs and tens of millions of GPU hours or even more. Searching for an optimal (or near-optimal) data mix is practically infeasible. Our paper focuses on offering a useful resource for continual pre-training in a massively multilingual scenario. To do this, we select the corpus to maintain diversity and balance among various languages and text types that align with the model's intended goal.

2.3 Model Training

We continue training the decoder-only Llama 3 models (8B parameters) using the causal language modeling objective, exposing the pre-trained model to new data and languages to develop our `EMMA-500` model. To enhance efficiency, we use training strategies that optimize memory usage, precision handling, and distributed training. `EMMA-500` Llama 3 series models are trained on the LUMI Supercomputer, powered by 100% renewable and carbon-neutral energy, utilizing 64 compute nodes with 256 AMD MI250x GPUs (512 AMD Graphics Compute Dies) with the GPT-NeoX framework (Andonian et al., 2023).

We continue training the base models in a full-parameter manner without modifying the tokenizer. The training setup includes a global batch size of 2048 and sequence lengths of 8192 tokens. Model- and data-specific settings are presented in Table 2. For training on the monolingual mix, the process spans 25,000 steps, accumulating a total of 419 billion Llama 3 tokens. For the bilingual mix that contains more tokens, we train for more steps, up to 40,000, leading to a total of 671 billion tokens. A key design choice in our study is the variation in training steps between the Monolingual and Bilingual mixes. We train each model for one epoch over its respective mix to ensure full exposure to the unique data distribution. Because the bilingual mix is inherently larger, this resulted in 15,000 additional steps for the bilingual models.

We employ the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.0001, betas set to [0.9, 0.95], and an epsilon of 1e-8. We experiment with different learning rates and evaluate early checkpoints trained up to 5,000 steps (84 billion tokens). We find that training with the original learning rate of 0.0003 used by the Llama 3 leads to instability, such as many fast spikes, resulting in poor performance. Thus, we opt for a smaller learning rate with a more stable training curve. All experiments consume more than 800k GPU hours on the LUMI Supercomputer. However, we could not perform a grid search on the learning rate due to the constraint of computing resources.

A cosine learning rate scheduler, with a warm-up of 1,000 and 2,000 iterations for monolingual and bilingual mixes, respectively, is used to regulate learning dynamics.

Base Model	Data Mix	Our Model Name	Steps	Warmup	Tokens
Llama 3	Monolingual (fig. 1b)	`EMMA-500` Llama 3 Mono	25,000	1,000	419B
	Bilingual (fig. 1a)	`EMMA-500` Llama 3 Bi	40,000	2,000	671B
Llama 3.1	Monolingual (fig. 1b)	`EMMA-500` Llama 3.1 Mono	25,000	1,000	419B
	Bilingual (fig. 1a)	`EMMA-500` Llama 3.1 Bi	40,000	2,000	671B

Table 2: Continual pre-trained models and settings.

3 Evaluation and Discussion

This section is structured to systematically evaluate the overall performance of our models on multilingual and bilingual benchmarks, assessing improvements in both text understanding and generation. We analyze the impact of bilingual continual pre-training on multilingual language models and conduct ablation studies to isolate the contributions of bilingual pre-training compared to monolingual training. We also focus on low-resource language performance, demonstrating how continual pre-training enhances representation and generalization for underrepresented languages. No single model can be universally the best among all baselines across the full spectrum of multilingual tasks, benchmarks, and languages. We provide an analysis of language gains and failure cases, identifying which languages benefit the most and highlighting remaining challenges.

Tasks and Benchmarks We evaluate all models on 7 tasks and 12 benchmarks that cover from 10 to 1500 languages. Table 9 shows the details of those tasks, benchmarks, evaluation metrics, the number of languages and samples per language, and the domain. We do not use LLMs-as-a-judge (Li et al., 2024a) for evaluation due to its well-known limitations, especially in multilingual scenarios, such as low agreement with human judges (de Wynter et al., 2024).

Baselines We consider open-weight decoder-only models with 7-9B parameters as baselines. We primarily compare our CPT models with the original Llama 3/3.1 base models (Dubey et al., 2024) and the LlaMAX models (Lu et al., 2024)

continually trained from Llama 3. We also compare with a wide set of baselines (Appendix C.3), including (1) Llama 2 models and their CPT models; (2) multilingual models, including recent advances such as Aya 23 (Aryabumi et al., 2024), Gemma (Team et al., 2024), Qwen (Yang et al., 2024), and Marco-LLM (Ming et al., 2024).

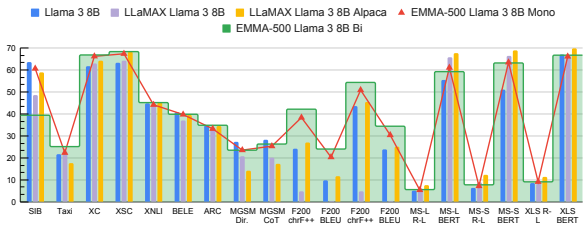
3.1 Impact of Mono- vs. Bilingual Training

To understand the role of continual training with bilingual translation data, we conduct controlled ablation studies by comparing monolingual continual pre-training (with the monolingual corpus only, i.e., data mix 2 in Figure 1b) vs. bilingual continual pre-training (with monolingual and bilingual extension, i.e., data mix 1 in Figure 1a). Figure 2 compares their average performance on each benchmark.

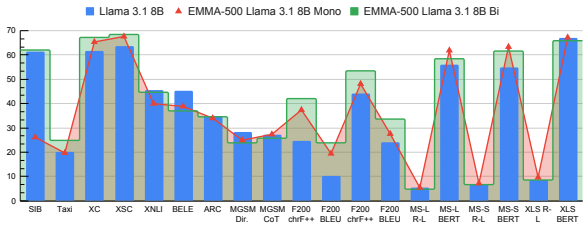
For CPT with Llama 3 as shown in Figure 2a, continual training on data mix with bilingual translation data consistently improves commonsense reasoning, natural language inference, reading comprehension, machine translation, and sometimes improves text classification when evaluated on Taxi1500, and retains similar summarization and math performance to the original base models.

For CPT with Llama 3.1 as shown in Figure 2b, continual training on data mix with bilingual translation data consistently improves text classification, commonsense reasoning, and machine translation. Except for summarization tasks, our CPT models trained on the monolingual mix are usually better than on the bilingual mix. More remarkably, CPT with both Llama 3 and 3.1 shows a large improvement on machine translation with an increase from 9% to 140% in terms of BLEU or chrF++ scores on translation directions from and to English on the Flores200 dataset.

Our study provided insights into how bilingual texts contribute to language adaptation, transferability, and performance stability. Overall, **continual training with bilingual translation data tends to improve the multilingual performance**, especially for machine translation, which benefits directly from the use of parallel texts in training.



(a) Llama 3



(b) Llama 3.1

Figure 2: Comparison of monolingual and bilingual CPT. The scores are averaged across all evaluated languages of the corresponding benchmarks. The baseline model LlaMAX does not have a CPT variant trained on Llama 3.1. Our models show a tendency for bilingual CPT to be better than monolingual CPT in most benchmarks and a remarkable advance on the Flores200 translation benchmark.

3.2 Low-Resource Language Performance

Low-resource languages often struggle with data scarcity and representation biases in multilingual models. We evaluate how baselines and our models perform on low-resource languages, and analyze whether continual pre-training on our data mixtures, especially the bilingual extension, can mitigate these limitations.

We focus on two benchmarks, SIB-200 and Flores200, with more than 200 languages, and one benchmark, Taxi1500, with more than 1500 languages, and examine the low-resource languages of these benchmarks according to the categorization in Table 7. Table 3 shows the average performance on low-resource languages. Our `EMMA-500` Llama 3 bilingual model obtained the best translation performance, followed by our Llama 3.1 bilingual one. They surpass the advanced Macro-LLM by a large margin on English to other translations and a small margin on other to English translations. For text classification, our models experience different levels of drops in the low-resource languages of SIB-200. Notably, our bilingual models perform the best on Taxi1500, showing their outstanding performance on low-resource languages validated on this massively multilingual benchmark. The results on low-resource languages show that **CPT on massive parallel data enhances the performance on low-resource languages**, especially for massively multilingual classification and translation.

Model	SIB		Flores200 (Eng-X)		Flores200 (X-Eng)	
	ACC	ACC	chrF++	BLEU	chrF++	BLEU
Llama 2 7B	22.74	18.04	16.23	4.93	31.26	13.69
Llama 2 7B Chat	26.02	16.01	18.12	5.30	32.61	13.05
CodeLlama 2 7B	23.84	17.15	15.98	4.57	29.47	11.52
LLaMAX Llama 2 7B	10.78	23.53	7.92	0.86	14.08	2.08
LLaMAX Llama 2 7B Alpaca	28.66	15.57	30.45	13.36	44.08	23.85
MaLA-500 Llama 2 10B v1	23.13	**24.76**	6.64	0.65	13.89	2.55
MaLA-500 Llama 2 10B v2	19.04	22.72	6.93	0.55	15.76	3.17
YaYi Llama 2 7B	25.03	17.98	16.04	4.81	32.30	13.76
TowerBase Llama 2 7B	19.53	18.06	17.15	5.10	32.44	14.50
TowerInstruct Llama 2 7B	20.81	17.92	16.84	3.54	26.32	5.24
EMMA-500 Llama 2 7B	**31.97**	21.73	**35.88**	**16.87**	**48.00**	**27.34**
Occiglot Mistral 7B v0.1	33.27	22.62	17.23	4.69	32.13	14.00
Occiglot Mistral 7B v0.1 Instruct	34.87	19.59	16.92	4.31	32.61	12.31
BLOOM 7B	17.73	14.93	12.32	2.67	28.51	9.96
BLOOMZ 7B	29.99	17.00	16.83	7.48	35.63	21.06
YaYi 7B	36.40	16.19	14.20	4.23	21.25	4.72
Aya 23 8B	42.20	22.52	16.89	6.41	33.31	14.48
Aya Expanse 8B	58.06	19.08	25.13	6.78	37.85	13.61
Gemma 7B	59.97	16.55	24.75	9.56	45.65	25.47
Gemma 2 9B	47.19	21.48	**28.50**	**12.71**	41.36	25.19
Qwen 1.5 7B	48.94	8.18	18.87	5.93	37.01	16.42
Qwen 2 7B	56.06	**23.01**	18.17	5.57	38.90	18.40
Qwen 2.5 7B	54.85	17.82	18.46	5.71	40.11	19.97
Marco-LLM GLO 7B	65.07	21.97	24.79	9.51	**45.89**	**26.38**
Llama 3 8B	**65.53**	**23.61**	25.84	10.51	45.72	25.49
Llama 3.1 8B	63.17	22.02	26.46	10.66	46.13	25.91
LLaMAX Llama 3 8B	49.70	22.91	4.93	0.49	4.87	0.52
LLaMAX Llama 3 8B Alpaca	60.91	20.48	**28.94**	**12.44**	**47.43**	**26.91**
EMMA-500 Llama 3 8B Mono	62.34	22.16	41.29	22.03	53.16	32.55
EMMA-500 Llama 3 8B Bi	39.65	**26.72**	**45.22**	**25.99**	**56.69**	**36.72**
EMMA-500 Llama 3.1 8B Mono	26.17	19.62	40.32	21.06	50.34	29.51
EMMA-500 Llama 3.1 8B Bi	**63.33**	25.42	45.07	25.80	55.84	35.91

Table 3: Performance on low-resource languages. The text classification task uses SIB-200 (SIB) and Taxi1500 (Taxi) datasets. **Underline and bold** represents the absolute best, underline means the second best, and **bold** signifies the best within a specific group. Our EMMA-500 models trained on bilingual data are the best two models in most cases.

3.3 Model Adaptability

It is unrealistic to expect a single model to perform the best across all tasks and benchmarks, given the inherent trade-offs in multilingual generalization, task specialization, and resource distribution. In practice, we experience some performance drop of our models on certain benchmarks. In this section, we analyze model adaptability by examining how CPT impacts performance when applied to different base models across a wide range of languages. To quantify the effect, we compute the performance gain as the difference between the CPT model and its corresponding base model on each benchmark.

We compare LLaMA 2, LLaMA 3, and LLaMA 3.1 as base models for continual pre-training, each offering progressively larger and more recent training corpora. LLaMA 2 was trained on 2T tokens, while LLaMA 3 significantly expands coverage with over 15T tokens. LLaMA 3.1 further updates the model with extensive data and long-context fine-tuning. This progression allows us to assess the adaptability of CPT across different model generations and data scales. Figure 3 shows the performance difference and the number of benchmarks where CPT models experience degradation.

CPT on Llama 2 observes a very small drop of BERTScore on MassiveSumm, where the score of the long subset decreases from 63.89 to 63.80 and the short one from 65.35 to 65.14. However, for Llama 3 and 3.1, both CPT on monolingual and bilingual mixes observe more cases of performance drops. For high-resource languages, Llama 3.1 degrades across more benchmarks than Llama 3. However, this trend diminishes notably when evaluating low-resource languages, where CPT remains effective to some extent.

This comparison provides insight into the effectiveness of continual pre-training with monolingual and bilingual data mixes in adapting different LLMs with various training to diverse linguistic settings. Our results align with the findings from Springer et al. (2025) that over-trained language models are harder to fine-tune. At a massively multilingual scale, we corroborate that **continuing pre-training on well-trained models**—particularly those already optimized for high-resource languages, e.g., Llama 3 and 3.1—**poses significant challenges when extending to hundreds of additional languages**.

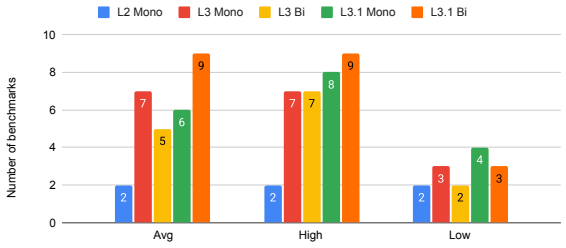


Figure 3: Model adaptability measured by the number of benchmarks on which CPT models are worse than the base model. CPT on LLaMA 2 (L2 Mono) shows a negligible BERTScore drop on MassiveSumm. More highly optimized models such as LLaMA 3 and 3.1 present greater challenges for effective continual pre-training compared to LLaMA 2, especially for high-resource languages, while the situation slightly eases for low-resource languages.

3.4 Per-Language Performance

Despite notable improvements, multilingual adaptation continues to present challenges. To scrutinize per-language performance, we evaluate each model by (1) counting the number of baseline models it outperforms across languages (from a total of 32 baselines; see Appendix C.3), and (2) calculating the percentage of languages where the model ranks

Models	BELEBELE			SIB-200			Taxi1500			Flores200		
	Top 3	Top 5	Top 10	Top 3	Top 5	Top 10	Top 3	Top 5	Top 10	Top 3	Top 5	Top 10
Llama 3 8B	0.82	0.82	25.41	43.41	65.37	95.61	7.17	13.80	38.16	0.00	0.00	25.12
LLaMAX Llama 3 8B	0.00	1.64	11.48	0.49	7.32	35.12	6.10	19.38	58.73	0.00	0.00	0.99
LLaMAX Llama 3 8B Alpaca	0.00	1.64	20.49	5.37	23.90	77.07	3.45	6.10	10.62	0.00	0.00	44.83
EMMA-500 Llama 3 8B Mono	2.46	9.02	37.70	30.24	43.41	76.10	14.60	22.50	44.79	1.97	77.34	96.06
EMMA-500 Llama 3 8B Bi	4.10	13.93	44.26	2.93	10.24	20.49	32.45	47.38	78.23	88.18	98.52	100.00

Table 4: Percentage of top-k rankings across four multilingual benchmarks. The darker color indicates the better. EMMA-500—especially with bilingual training—consistently outperforms baselines on Taxi1500 and Flores200. While Llama 3 and LlaMAX models yield higher accuracy on BELEBELE, EMMA-500 models cover more languages with improved relative performance.

among the top k performers. This provides a fine-grained view of cross-lingual competitiveness.

Table 4 presents the top-k performance percentage across four multilingual benchmarks—BELEBELE, SIB-200, Taxi1500, and FLORES-200. The results demonstrate that EMMA-500 models, particularly those trained with bilingual data, achieve higher top-k percentages on low-resource-heavy benchmarks like Taxi-1500 and FLORES-200. In contrast, base LLaMA 3 and LLaMAX variants exhibit more limited top-tier rankings. Notably, for BELEBELE, a machine comprehension benchmark, Llama 3 and 3.1 are not very strong models on this benchmark as evaluated in Appendix D.7. Existing CPT models, like LlaMAX Llama 3, and our CPT models obtained degraded performance. Despite having a decreased average accuracy, **our models have better performance for more languages** than the base model and the LlaMAX Llama 3 model.

We further look at the evaluation results of selected languages on BELEBELE in Table 5 for a mix of low-resource (e.g., Khalkha Mongolian, Northern Sotho, Plateau Malagasy) and high-resource (e.g., English, French, Italian) languages. Table 5 reports absolute scores and the delta (Δ) between each CPT model and the Llama 3 8B base model. Results show that while CPT variants generally improve scores for low-resource languages—especially in the EMMA-500 setups—they tend to suffer from notable performance degradation in high-resource languages. However, this trade-off is arguably acceptable as the main focus of this paper is to adapt LLMs into low-resource languages.

These analyses provide a deeper understanding of the strengths and limitations of massively multilingual adaptation and guide future improvements in multilingual language model training and evaluation. First, they show averaging scores across languages to compare model performance, while convenient, has several important limitations, e.g., obscuring disparities in performance across languages and resources, and ignoring per-language variance. Second, this comparison of per-language performance highlights the benefits of our CPT models in promoting cross-lingual competitiveness.

Model	Average	khk_Cyrl	nso_Latn	plt_Latn	eng_Latn	fra_Latn	ita_Latn
Llama 3 8B	40.73	35.11	29.22	30.44	74.56	60.89	60.22
LLaMAX	36.96	34.67	30.33	34.11	62.33	48.78	47.33
Δ	-3.77	-0.44	1.11	3.67	-12.23	-12.11	-12.89
LLaMAX Alpaca	39.41	35.56	30.78	33.78	65.78	54.67	53.67
Δ	-1.32	0.45	1.56	3.34	-8.78	-6.22	-6.55
EMMA-500 Mono	39.73	38.44	37.00	34.78	56.00	51.11	49.56
Δ	-1.00	3.33	7.78	4.34	-18.56	-9.78	-10.66
EMMA-500 Bi	39.84	39.33	35.22	33.00	56.78	49.67	46.00
Δ	-0.89	4.22	6.00	2.56	-17.78	-11.22	-14.22

Table 5: Performance of selected languages in the BELEBELE benchmark. Δ denotes the difference between the CPT model's score and that of the base model. All models are based on Llama 3 8B. CPT on Llama 3 exhibits a significant drop in some high-resource languages, much bigger than the increase in some low-resource languages.

3.5 Overall Results

We evaluate the model on a comprehensive set of multilingual and bilingual benchmarks, assessing both language understanding and generation tasks. Table 10 in Appendix D.1 and Table 11 in Appendix D.1 present the results of deterministic and generation tasks. Our EMMA-500 models are the best at machine translation (Flores200) and competitive at text classification (Taxi1500 and SIB-200) and commonsense reasoning (XCOPA and XStoryCloze). While our models are competitive in many cases, we also observe some performance degradation, such asin math tasks Performance inevitably varies across tasks and languages, reflecting trade-offs in model capacity and data representation. Appendix D presents a deeper dive into the performance details across all tasks and languages evaluated, providing insights into model behavior across multilingual datasets and benchmarks. We also provide detailed per-language re-

sults, available at https://mala-lm.github.io/emma-500-gen2

4 Conclusion

This work advances massively multilingual adaptation of LLMs using bilingual translation data across 2,500+ language pairs (500+ languages). The four released EMMA-500 models using Llama 3 and 3.1 trained on both monolingual and bilingual data mixes, including the newly compiled MaLA translation corpus, establish new benchmarks in multilingual coverage while maintaining competitive performance across 7 diverse tasks. Notably, they achieve state-of-the-art results on machine translation while showing robust generalization to text classification and reasoning tasks. Despite advancements, achieving consistently high performance across diverse benchmarks is constrained by linguistic and task variability, which highlights intrinsic tensions between scale and specialization in multilingual NLP.

Acknowledgment

This work has received funding from the HPLT project, supported by the European Union's Horizon Europe research and innovation programme (Grant No. 101070350) and by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No. 10052546), and from the Digital Europe Programme (Grant No. 101195233, OpenEuroLLM). The authors wish to acknowledge CSC – IT Center for Science, Finland, for computational resources. We thank Indraneil Paul for his contributions to preparing the code data.

Limitations

Multilingual Benchmark Multilingual language models are designed to accommodate users across various linguistic and cultural backgrounds. However, many widely used multilingual evaluation benchmarks, including some in this research, are developed through human or machine translation. As a result, they often emphasize knowledge and subject matter primarily from English-speaking sources, with potential translation-related distortions that can undermine the accuracy of model assessments (Chen et al., 2024). This discrepancy highlights the need for more comprehensive, native-language test sets that better reflect the full range of linguistic diversity. We encourage collaborative efforts to create large-scale benchmarks that offer a more reliable evaluation of these models across different languages and cultures.

Human Evaluation Although human assessment is a valuable tool, it presents challenges such as variability, subjectivity, and high costs—especially when evaluating models across numerous languages. Recruiting expert annotators proficient in less common languages is particularly difficult, making large-scale human evaluation unfeasible. Even assessing a limited selection of languages requires significant resources. While this study acknowledges these constraints, we recognize that human evaluations play a crucial role in supplementing automated assessment methods. Given these limitations, we rely on automatic evaluation tools to ensure scalability and consistency, despite their imperfections.

Model Performance Despite its strengths in multilingual processing, our models face challenges in areas such as mathematical reasoning and machine reading comprehension. Its performance on machine-translated mathematical benchmarks remains limited, likely due to the inherent difficulty of numerical reasoning and translation artifacts that may obscure problem clarity. Similarly, while our model achieves improvements in various NLP tasks, it struggles with reading comprehension, which demands deep contextual understanding and logical inference. Addressing these weaknesses will require further refinements, such as incorporating domain-specific training data or exploring alternative model architectures optimized for these challenges. Besides, averaging scores across languages can mask important disparities, such as strong performance on high-resource languages and poor results on low-resource ones. It also ignores linguistic diversity and per-language difficulty, potentially leading to misleading conclusions about overall model performance.

Real-world Usage This research primarily focuses on enhancing continual training with bilingual texts and improving language model performance through continual pre-training. However, the model is not yet suitable for deployment in real-world applications. It has not undergone thorough human alignment processes or adversarial robustness testing (red-teaming) to ensure safety and reliability. While our work contributes to advancements in multilingual NLP, additional refinements—such

as aligning the model with human preferences and conducting rigorous safety evaluations—are necessary before it can be practically implemented.

Data Mix We examine two manually constructed data mixes to assess the impact of CPT with bilingual translation data. While numerous alternative configurations are possible, systematically validating the effectiveness of different data mixes or searching for the optimal data mixes requires a lot of computing resources, which are not affordable for a small research team like us. Nonetheless, the two data mixes carefully decided by us already demonstrate measurable improvements in the downstream evaluation, which highlight the utility of our design choices.

Model Training Model training involves a few hyperparameters, and a wide range of base models can be selected for continual pre-training. When preparing the bilingual translation data for training, the number of lines is intuitively chosen. However, an exhaustive search over all possible configuration combinations—including data volume, model choice, and hyperparameter settings—would incur prohibitive computational costs. Consequently, we focus on a limited set of configurations that are feasible under available resources.

Community Collaboration This study was conducted without direct community collaboration. Nonetheless, we recognize the value of community collaboration and are open to future partnerships with researchers and practitioners in this area.

References

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, and 1 others. 2022. A few thousand translations go a long way! leveraging pre-trained models for african news translation. *arXiv preprint arXiv:2205.02022*.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2023. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *CoRR*, abs/2309.07445.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Jason Phang, Shivanshu Purohit, Hailey Schoelkopf, Dashiell Stander, Tri Songz, Curt Tigges, Benjamin Thérien, and 2 others. 2023. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch.

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*.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The BELEBELE benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. 2024. Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models? *arXiv preprint arXiv:2406.12822*.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

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*, abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. 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*. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and 1 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Adrian de Wynter, Ishaan Watts, Tua Wongsangaroonsri, Minghui Zhang, Noura Farra, Nektar Ege Altıntoprak, Lena Baur, Samantha Claudet, Pavel Gajdusek, Can Gören, Qilong Gu, Anna Kaminska, Tomasz Kaminski, Ruby Kuo, Akiko Kyuba, Jongho Lee, Kartik Mathur, Petter Merok, Ivana Milovanović, and 14 others. 2024. RTP-LX: Can LLMs evaluate toxicity in multilingual scenarios? *arXiv preprint 2404.14397*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

EdTeKLA Research Group. 2022. Indigenous languages corpora. `https://github.com/EdTeKLA/IndigenousLanguages_Corpora`. Accessed: 2024-05-30.

Manuel Faysse. 2023. Dataset card for "project gutenberg".

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *arXiv preprint 2404.17790*.

Kazuki Fujii, Yukito Tajima, Sakae Mizuki, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Masanari Ohi, Masaki Kawamura, Taishi Nakamura, Takumi Okamoto, Shigeki Ishida, Kakeru Hattori, Youmi Ma, Hiroya Takamura, Rio Yokota, and Naoaki Okazaki. 2025. Rewriting pre-training data boosts llm performance in math and code. *Preprint*, arXiv:2505.02881.

Ana-Paula Galarreta, Andrés Melgar, and Arturo Oncevay. 2017. Corpus creation and initial SMT experiments between Spanish and Shipibo-konibo. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 238–244, Varna, Bulgaria. INCOMA Ltd.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, and 1 others. 2023. A framework for few-shot language model evaluation. Zenodo.

Yvette Gbedevi. 2019. Ewe language corpus. Accessed: 2024-08-27.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

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*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhen Leng Thai, Kai Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, and 6 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *CoRR*, abs/2404.06395.

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.

Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, Jinan Xu, Jian-Yun Nie, and Yang Liu. 2025. A survey on large language

models with multilingualism: Recent advances and new frontiers. *Preprint*, arXiv:2405.10936.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and 1 others. 2024a. Emma-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint arXiv:2409.17892*.

Shaoxiong Ji, Timothee Mickus, Vincent Segonne, and Jörg Tiedemann. 2024b. Can machine translation bridge multilingual pretraining and cross-lingual transfer learning? In *Proceedings of LREC-COLING*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France. European Language Resources Association.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Yacine Jernite, Margaret Mitchell, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Dzmitry Bahdanau, Leandro von Werra, and Harm de Vries. 2023. The stack: 3 TB of permissively licensed source code. *Trans. Mach. Learn. Res.*, 2023.

Minato Kondo, Takehito Utsuro, and Masaaki Nagata. 2024. Enhancing translation accuracy of large language models through continual pre-training on parallel data. *arXiv preprint 2407.03145*.

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

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8186–8213, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Mike Lewis. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Yudong Li, Yuqing Zhang, Zhe Zhao, Linlin Shen, Weijie Liu, Weiquan Mao, and Hui Zhang. 2022. CSL: A large-scale chinese scientific literature dataset. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3917–3923.

Zihao Li, Shaoxiong Ji, Hengyu Luo, and Jörg Tiedemann. 2025. Rethinking multilingual continual pretraining: Data mixing for adapting llms across languages and resources. *arXiv preprint 2504.04152*.

Zihao Li, Shaoxiong Ji, Timothee Mickus, Vincent Segonne, and Jörg Tiedemann. 2024b. A comparison of language modeling and translation as multilingual pretraining objectives. In *Proceedings of EMNLP*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. MaLA-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, and 1 others. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, and 1 others. 2024. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*.

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. *arXiv preprint arXiv:2407.05975*.

Risto Luukkonen, Jonathan Burdge, Elaine Zosa, Aarne Talman, Ville Komulainen, Väinö Hatanpää, Peter Sarlin, and Sampo Pyysalo. 2024. Poro 34b and the blessing of multilinguality. *arXiv preprint 2404.01856*.

Chunlan Ma, Ayyoob ImaniGooghari, Haotian Ye, Ehsaneddin Asgari, and Hinrich Schütze. 2023. Taxi1500: A multilingual dataset for text classification in 1500 languages. *Preprint*, arXiv:2305.08487.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018. Probabilistic finite-state morphological segmenter for wixarika (huichol) language. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Pratyush Maini, Skyler Seto, Richard He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 14044–14072. Association for Computational Linguistics.

Masakhane. 2022. lacuna_pos_ner: POS and NER for african languages. `https://github.com/masakhane-io/lacuna_pos_ner`. Accessed: 2024-05-30.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 3158–3163. European Language Resources Association (ELRA).

Lingfeng Ming, Bo Zeng, Chenyang Lyu, Tianqi Shi, Yu Zhao, Xue Yang, Yefeng Liu, Yiyu Wang, Linlong Xu, Yangyang Liu, Xiaohu Zhao, Hao Wang, Heng Liu, Hao Zhou, Huifeng Yin, Zifu Shang, Haijun Li, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-llm: Bridging languages via massive multilingual training for cross-lingual enhancement. *Preprint*, arXiv:2412.04003.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, and 1 others. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Jonathan Mukiibi, Babirye Claire, and Nakatumba-Nabende Joyce. 2021. An english-luganda parallel corpus.

NLLB Team, Marta R, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

John E Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2020. Overcoming resistance: The normalization of an Amazonian tribal language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–13, Suzhou, China. 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 *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Indraneil Paul, Goran Glavas, and Iryna Gurevych. 2024. Ircoder: Intermediate representations make language models robust multilingual code generators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15023–15041. Association for Computational Linguistics.

Indraneil Paul, Haoyi Yang, Goran Glavas, Kristian Kersting, and Iryna Gurevych. 2025. Obscuracoder: Powering efficient code LM pre-training via obfuscation grounding. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Jackson Petty, Sjoerd van Steenkiste, and Tal Linzen. 2024. How does code pretraining affect language model task performance? *CoRR*, abs/2409.04556.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ruchir Puri, David S. Kung, Geert Janssen, Wei Zhang, Giacomo Domeniconi, Vladimir Zolotov, Julian Dolby, Jie Chen, Mihir R. Choudhury, Lindsey Decker, Veronika Thost, Luca Buratti, Saurabh Pujar, Shyam Ramji, Ulrich Finkler, Susan Malaika, and Frederick Reiss. 2021. Codenet: A large-scale AI for code dataset for learning a diversity of coding tasks.

In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, and 1 others. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwaraknath Gnaneshwar, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. 2025. Procedural knowledge in pretraining drives reasoning in large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and 1 others. 2022. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint*.

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.

Archchana Sindhujan, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. When llms struggle: Reference-less translation evaluation for low-resource languages. *Preprint*, arXiv:2501.04473.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Raghavi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15725–15788. Association for Computational Linguistics.

Luca Soldaini and Kyle Lo. 2023. peS2o (Pretraining Efficiently on S2ORC) Dataset. Technical report, Allen Institute for AI. ODC-By, https://github.com/allenai/pes2o.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. 2025. Overtrained language models are harder to fine-tune. *arXiv preprint arXiv:2503.19206*.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. *CoRR*, abs/2412.02595.

Marc Szafraniec, Baptiste Rozière, Hugh Leather, Patrick Labatut, François Charton, and Gabriel Synnaeve. 2023. Code translation with compiler representations. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*.

Carnegie Mellon University. 2010. Haitian creole data. Accessed: 2024-05-30.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

José Mateo Lino Cajero Velázquez. 2021. Spanish-hnähnü corpus.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model. *arXiv preprint 2307.06018*.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint 2309.11674*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2023. Lego-MT: Learning detachable models for massively multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11518–11533, Toronto, Canada. Association for Computational Linguistics.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2020. Chren: Cherokee-english machine translation for endangered language revitalization. In *EMNLP2020*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Details of MaLA Translation Corpus

A.1 Data Sources

Table 6 lists the corpora and collections we use as bilingual data sources in this work. A data source being bilingual means that it contains corresponding, or parallel, text records in two languages. This usually means one side has been translated, either by a machine or a human. Pedantically, massively parallel bilingual corpora can be considered a special case of multilingual corpora containing sentences in different languages that share the same meaning. In this work, some data sources have parallel text data in more than two languages, but in such cases, we always iteratively determine one as the source and one as the target language, i.e., create bilingual corpora out of them. In some source datasets, the source language is not explicitly stated, and as such we pick one language to regard as the source language without actual knowledge of the direction of translation. In such cases, we tend to pick a high-resource language, such as English, as the source language.

Metadata In the case of bilingual corpora, we define each record in the `JSONL` file to consist of the fields `src_lang`, `src_txt`, `tgt_lang`, `tgt_txt`, `url`, `collection`, `source`, `original_src_lang`, and `original_tgt_lang`. The contents of these fields are as follows. The fields `src_txt` and `tgt_txt` contain the parallel text data in the source and target language, respectively, and `src_lang` and `tgt_lang` are the ISO 639-3 language codes of those languages. Identically with the monolingual corpora, `url` contains a URL indicating the web address from which the text data has been extracted, if available, `collection` the name of the collection, and `source` the name of a more specific part of the collection which the text data was extracted from. Lastly, `original_src_lang` and `original_tgt_lang` contain the language denotations of the source and target language, respectively, as they are given in the source data.

A.2 Supported Languages

Table 7 shows the language codes of the MaLA corpus (Ji et al., 2024a), where "unseen" means the languages are not used for training `EMMA-500`. The classification system for token counts categorizes language resources based on their size into five distinct tiers: "high" for resources exceeding 1 billion tokens, indicating a vast amount of data; "medium-high" for those with more than 100 million tokens, reflecting a substantial dataset; "medium" for resources that contain over 10 million tokens, representing a moderate size; "medium-low" for datasets with over "1 million tokens", indicating a smaller yet significant amount of data; and finally, "low" for resources containing less than 1 million tokens, which suggests a minimal data presence. This hierarchy helps in understanding the scale and potential utility of the language resources available.

A.3 Token Counts

Figure 4 shows the numbers of segments and tokens across all language pairs in the `MaLA` translation corpus.

B Details of Data Mixes

Table 8 provides how different data types and resource categories contribute to the overall dataset composition. Our data mixes aim to make a balanced distribution over high- and low-resources. However, due to the nature of high-resource languages, they still contribute to a large portion. For example, English research papers form a substantial portion. Medium-high and medium-resource monolingual and bilingual texts contribute a lot. Medium and low-resource languages consist of 33% and 19% of bilingual and monolingual mixes, respectively.

B.1 Additional Code Data

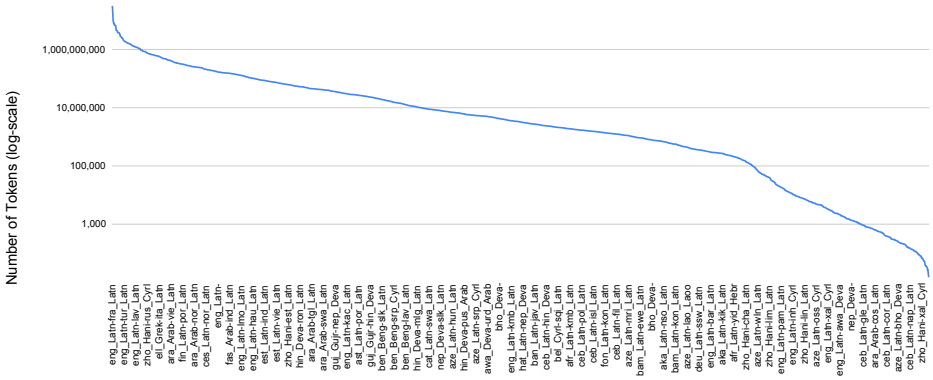
Code Data We source code data from a deduplicated version of The Stack (Kocetkov et al., 2023) and oversample files from the algorithmic coding (Puri et al., 2021) and data science domains [11] owing to the benefits of training models on self-contained code (Fujii et al., 2025). We discard all non-data programming languages that occur fewer than 50k times, with the exception of `llvm`, following prior work detailing its importance in multi-lingual and low-resource code generation (Szafraniec et al., 2023; Paul et al., 2024). We also discard samples that manifest code in rare but valid extensions. Finally, we source from data-heavy formats but follow precedent (Lozhkov et al., 2024) and subsample them more aggressively. The surviving data is filtered as follows:

- For files forked more than 25 times, we retain them if the average line length is less than 120,

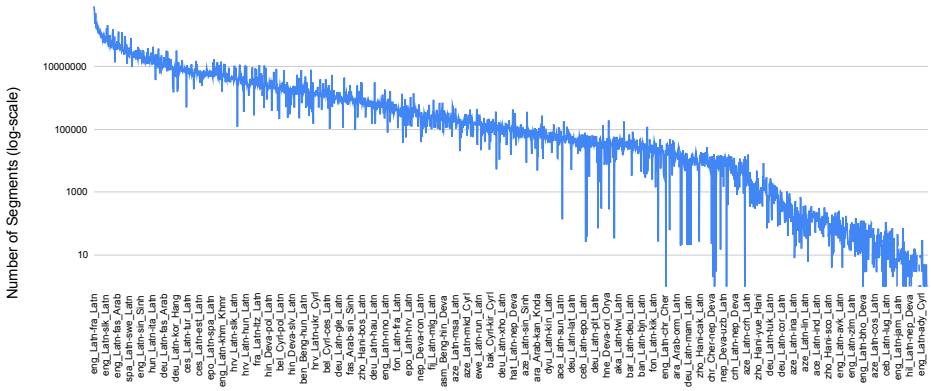
[11] https://huggingface.co/datasets/AlgorithmicResearchGroup/arxiv_research_code

Dataset	Languages	URL
CMU Haitian Creole (University, 2010)	Haitian Creole	http://www.speech.cs.cmu.edu/haitian/
English-Luganda Parallel Corpus (Mukiibi et al., 2021)	Luganda-English	https://zenodo.org/records/4764039
Ewe language corpus (Gbedevi, 2019)	Ewe—English	https://www.kaggle.com/datasets/yvicherita/ewe-language-corpus
Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 (Joanis et al., 2020)	Inuktitut–English	https://nrc-digital-repository.canada.ca/eng/view/object/?id=c7e34fa7-7629-43c2-bd6d-19b32bf64f60
AmericasNLP 2021 (Gutierrez-Vasques et al., 2016; Velázquez, 2021; Galarreta et al., 2017)	10	https://turing.iimas.unam.mx/americasnlp/americasnlp_2021.html; https://github.com/AmericasNLP/americasnlp2021
AmericasNLP 2023 (Mager et al., 2018; Gutierrez-Vasques et al., 2016; Chiruzzo et al., 2020; Ortega et al., 2020; Velázquez, 2021)	11	https://turing.iimas.unam.mx/americasnlp/2023_st.html
AmericasNLP 2022	4	https://github.com/AmericasNLP/americasnlp2022
Indigenous Languages Corpora (EdTeKLA Research Group, 2022)	Cree	https://github.com/EdTeKLA/IndigenousLanguages_Corpora
ACES (Amrhein et al., 2022)	146	https://huggingface.co/datasets/nikitam/ACES
ChrEn (Zhang et al., 2020)	2	https://huggingface.co/datasets/chr_en
NusaX-MT (Winata et al., 2023)	12	https://huggingface.co/datasets/indonlp/NusaX-MT
Lego-MT (Yuan et al., 2023)	433	https://github.com/CONE-MT/Lego-MT
Tatoeba (Tiedemann, 2020)	487	https://github.com/Helsinki-NLP/Tatoeba-Challenge/tree/master/data
NLLB (NLLB Team et al., 2022)	202	https://huggingface.co/datasets/allenai/nllb
Lacuna Project (Masakhane, 2022)	1	https://github.com/masakhane-io/lacuna_pos_ner
lafand-mt (Adelani et al., 2022)	21	https://github.com/masakhane-io/lafand-mt/tree/main

Table 6: Source datasets used for compiling the MaLA bilingual translation data.



(a) Number of tokens



(b) Number of segments

Figure 4: Numbers of segments and tokens across all language pairs in the MaLA bilingual translation corpus.

the maximum line length is less than 300, and the alphanumeric fraction is more than 30%.

- For files forked between 15 and 25 times, we retain them if the average line length is less than 90, the maximum line length is less than 150, and the alphanumeric fraction is more than 40%.
- For files forked less than 15 times, we retain

Category	Languages	Language Codes
high	27	fra_Latn, mon_Cyrl, kat_Geor, tgk_Cyrl, kaz_Cyrl, glg_Latn, hbs_Latn, kan_Knda, mal_Mlym, rus_Cyrl, cat_Latn, hye_Armn, guj_Gujr, slv_Latn, fil_Latn, bel_Cyrl, isl_Latn, nep_Deva, mlt_Latn, pan_Guru, afr_Latn, urd_Arab, mkd_Cyrl, aze_Latn, deu_Latn, eng_Latn, ind_Latn
low	210	prs_Arab, nqo_Nkoo, emp_Latn, pfl_Latn, teo_Latn, gpe_Latn, izz_Latn, shn_Mymr, hak_Latn, pls_Latn, evn_Cyrl, djk_Latn, toj_Latn, nog_Cyrl, ctu_Latn, tca_Latn, jiv_Latn, ach_Latn, mrj_Latn, ajp_Arab, apc_Arab, tab_Cyrl, hvn_Latn, tls_Latn, bak_Latn, ndc_Latn, trv_Latn, top_Latn, kjh_Cyrl, guh_Latn, mni_Mtei, csy_Latn, noa_Latn, dov_Latn, bho_Deva, kon_Latn, hne_Deva, kcg_Latn, mni_Beng, hus_Latn, pau_Latn, jbo_Latn, dtp_Latn, kmb_Latn, hau_Arab, pdc_Latn, nch_Latn, acf_Latn, bim_Latn, ixl_Latn, dty_Deva, kas_Arab, lrc_Arab, alz_Latn, lez_Cyrl, lld_Latn, tdt_Latn, acm_Arab, bih_Deva, mzh_Latn, guw_Latn, rop_Latn, rwo_Latn, ahk_Latn, qub_Latn, kri_Latn, gub_Latn, laj_Latn, sxn_Latn, luo_Latn, tly_Latn, pwn_Latn, mag_Deva, xav_Latn, bum_Latn, ubu_Latn, roa_Latn, mah_Latn, tsg_Latn, gcr_Latn, arn_Latn, csb_Latn, guc_Latn, bat_Latn, knj_Latn, cre_Latn, bus_Latn, anp_Deva, aln_Latn, nah_Latn, zai_Latn, kpv_Cyrl, enq_Latn, gvl_Latn, wal_Latn, fiu_Latn, swh_Latn, crh_Latn, nia_Latn, bqc_Latn, map_Latn, atj_Latn, npi_Deva, bru_Latn, din_Latn, pis_Latn, gur_Latn, cuk_Latn, zne_Latn, cdo_Latn, lhu_Latn, pcd_Latn, mas_Latn, bis_Latn, ncj_Latn, ibb_Latn, tay_Latn, bts_Latn, tzj_Latn, bzj_Latn, cce_Latn, jvn_Latn, ndo_Latn, rug_Latn, koi_Cyrl, mco_Latn, fat_Latn, olo_Latn, inb_Latn, mkn_Latn, qvi_Latn, mak_Latn, ktu_Latn, nrm_Latn, kua_Latn, san_Latn, nbl_Latn, kik_Latn, dyu_Latn, sgs_Latn, msm_Latn, mnw_Latn, zha_Latn, sja_Latn, xal_Cyrl, rmc_Latn, ami_Latn, sda_Latn, tdx_Latn, yap_Latn, tzh_Latn, sus_Latn, ikk_Latn, bas_Latn, nde_Latn, dsb_Latn, seh_Latn, knv_Latn, amu_Latn, dwr_Latn, iku_Cans, uig_Latn, bxr_Cyrl, tcy_Knda, mau_Latn, aoj_Latn, gor_Latn, cha_Latn, fip_Latn, chr_Cher, mdf_Cyrl, arb_Arab, quw_Latn, shp_Latn, spp_Latn, frp_Latn, ape_Latn, cbk_Latn, mnw_Mymr, mfe_Latn, jam_Latn, lad_Latn, awa_Deva, mad_Latn, ote_Latn, shi_Latn, btx_Latn, maz_Latn, ppk_Latn, smn_Latn, twu_Latn, blk_Mymr, msi_Latn, naq_Latn, tly_Arab, wuu_Hani, mos_Latn, cab_Latn, zlm_Latn, gag_Latn, suz_Deva, ksw_Mymr, gug_Latn, nij_Latn, nov_Latn, srm_Latn, jac_Latn, nyu_Latn, yom_Latn, gui_Latn
medium	68	tha_Thai, kat_Latn, lim_Latn, tgk_Arab, che_Cyrl, lav_Latn, xho_Latn, war_Latn, nan_Latn, grc_Grek, orm_Latn, zsm_Latn, cnh_Latn, yor_Latn, arg_Latn, tgk_Latn, azj_Latn, tel_Latn, slk_Latn, pap_Latn, zho_Hani, sme_Latn, tgl_Latn, uzn_Cyrl, als_Latn, san_Deva, azb_Arab, ory_Orya, lmo_Latn, bre_Latn, mvf_Mong, fao_Latn, oci_Latn, sah_Cyrl, sco_Latn, tuk_Latn, aze_Arab, hin_Deva, haw_Latn, glk_Arab, oss_Cyrl, lug_Latn, tet_Latn, tsn_Latn, hrv_Latn, gsw_Latn, arz_Arab, vec_Latn, mon_Latn, ilo_Latn, ctd_Latn, ben_Beng, roh_Latn, kal_Latn, asm_Beng, srp_Latn, bod_Tibt, hif_Latn, rus_Latn, nds_Latn, lus_Latn, ido_Latn, lao_Laoo, tir_Ethi, chv_Cyrl, wln_Latn, kaa_Latn, pnb_Arab
medium-high	79	div_Thaa, som_Latn, jpn_Japn, hat_Latn, sna_Latn, heb_Hebr, bak_Cyrl, nld_Latn, tel_Telu, kin_Latn, msa_Latn, gla_Latn, bos_Latn, dan_Latn, smo_Latn, ita_Latn, mar_Deva, pus_Arab, srp_Cyrl, spa_Latn, lat_Latn, hmn_Latn, sin_Sinh, zul_Latn, bul_Cyrl, amh_Ethi, ron_Latn, tam_Taml, khm_Khmr, nno_Latn, cos_Latn, fin_Latn, ori_Orya, uig_Arab, hbs_Cyrl, gle_Latn, cym_Latn, vie_Latn, kor_Hang, lit_Latn, yid_Hebr, ara_Arab, sqi_Latn, pol_Latn, tur_Latn, swa_Latn, hau_Latn, ceb_Latn, eus_Latn, kir_Cyrl, mlg_Latn, jav_Latn, snd_Arab, sot_Latn, por_Latn, uzb_Cyrl, fas_Arab, nor_Latn, est_Latn, hun_Latn, ibo_Latn, ltz_Latn, swe_Latn, tat_Cyrl, ast_Latn, mya_Mymr, uzb_Latn, sun_Latn, ell_Grek, ces_Latn, mri_Latn, ckb_Arab, kur_Latn, kaa_Cyrl, nob_Latn, ukr_Cyrl, fry_Latn, epo_Latn, nya_Latn
medium-low	162	aym_Latn, rue_Cyrl, rom_Latn, dzo_Tibt, poh_Latn, sat_Olck, ary_Arab, fur_Latn, mbt_Latn, bpy_Beng, iso_Latn, pon_Latn, glv_Latn, new_Deva, gym_Latn, bgp_Latn, kac_Latn, abt_Latn, quc_Latn, otq_Latn, sag_Latn, cak_Latn, avk_Latn, pam_Latn, meo_Latn, tum_Latn, bam_Latn, kha_Latn, syr_Syrc, kom_Cyrl, nhe_Latn, bal_Arab, srd_Latn, krc_Cyrl, lfn_Latn, bar_Latn, rcf_Latn, nav_Latn, nnb_Latn, sdh_Arab, aka_Latn, bew_Cyrl, bbc_Latn, meu_Latn, zza_Latn, ext_Latn, yue_Hani, ekk_Latn, xmf_Geor, nap_Latn, mzn_Arab, pcm_Latn, lij_Latn, myv_Cyrl, scn_Latn, dag_Latn, ban_Latn, twi_Latn, udm_Cyrl, som_Arab, nso_Latn, pck_Latn, crs_Latn, acr_Latn, tat_Latn, afb_Arab, uzs_Arab, hil_Latn, mgh_Latn, tpi_Latn, ady_Cyrl, pag_Latn, kiu_Latn, ber_Latn, iba_Latn, ksh_Latn, plt_Latn, lin_Latn, chk_Latn, tzo_Latn, tlh_Latn, ile_Latn, lub_Latn, hui_Latn, min_Latn, bjn_Latn, szl_Latn, kbp_Latn, inh_Cyrl, que_Latn, ven_Latn, vls_Latn, kbd_Cyrl, run_Latn, wol_Latn, ace_Latn, ada_Latn, kek_Latn, yua_Latn, tbz_Latn, gom_Latn, ful_Latn, mrj_Cyrl, abk_Cyrl, tuc_Latn, stq_Latn, mwl_Latn, tvl_Latn, quh_Latn, gom_Deva, mhr_Cyrl, fij_Latn, grn_Latn, zap_Latn, mam_Latn, mps_Latn, tiv_Latn, ksd_Latn, ton_Latn, bik_Latn, vol_Latn, ava_Cyrl, tso_Latn, szy_Latn, ngu_Latn, hyw_Armn, fon_Latn, skr_Arab, kos_Latn, tyz_Latn, kur_Arab, srn_Latn, tyv_Cyrl, bci_Latn, vep_Latn, crh_Cyrl, kpg_Latn, hsb_Latn, ssw_Latn, zea_Latn, ewe_Latn, ium_Latn, diq_Latn, ltg_Latn, nzi_Latn, guj_Deva, ina_Latn, pms_Latn, bua_Cyrl, lvs_Latn, eml_Latn, hmo_Latn, kum_Cyrl, kab_Latn, chm_Cyrl, cor_Latn, cfm_Latn, alt_Cyrl, bcl_Latn, ang_Latn, frr_Latn, mai_Deva
unseen	393	rap_Latn, pmf_Latn, lsi_Latn, dje_Latn, bkx_Latn, ipk_Latn, syw_Deva, ann_Latn, bag_Latn, bat_Cyrl, chu_Cyrl, gwc_Arab, adh_Latn, szy_Hani, shi_Arab, njy_Latn, pdu_Latn, buo_Latn, cuv_Latn, udg_Mlym, bax_Latn, tio_Latn, kjb_Latn, taj_Deva, lez_Latn, olo_Cyrl, rnl_Latn, bri_Latn, inh_Latn, kas_Cyrl, wni_Latn, anp_Latn, tsc_Latn, mgg_Latn, udi_Cyrl, mdf_Latn, agr_Latn, xty_Latn, llg_Latn, nge_Latn, gan_Latn, tuv_Latn, stk_Latn, nut_Latn, thy_Thai, lgr_Latn, hnj_Latn, dar_Cyrl, aia_Latn, lwl_Thai, tnl_Latn, tvs_Latn, jra_Khmr, tay_Hani, gal_Latn, ybi_Deva, snk_Arab, gag_Cyrl, tuk_Cyrl, trv_Hani, ydd_Hebr, kea_Latn, gbm_Deva, kwi_Latn, hro_Latn, rki_Latn, quy_Latn, tdg_Deva, zha_Hani, pcg_Mlym, tom_Latn, nsn_Latn, quf_Latn, jmx_Latn, kqr_Latn, mrn_Latn, bxa_Latn, abc_Latn, mve_Arab, lfa_Latn, qup_Latn, yin_Latn, roo_Latn, mrw_Latn, nxa_Latn, yrk_Cyrl, bem_Latn, kvt_Latn, csw_Cans, bjr_Latn, mgm_Latn, ngn_Latn, pib_Latn, quz_Latn, awb_Latn, myk_Latn, otq_Arab, ino_Latn, tkd_Latn, bef_Latn, bug_Bugi, aeu_Latn, nlv_Latn, dty_Latn, bkc_Latn, mmu_Latn, hak_Hani, sea_Latn, mlk_Latn, cbr_Latn, lmp_Latn, tnn_Latn, qvz_Latn, pbt_Arab, cjs_Cyrl, mlw_Latn, mnf_Latn, bfm_Latn, dig_Latn, thk_Latn, zxx_Latn, lkb_Latn, chr_Latn, pnt_Latn, vif_Latn, fli_Latn, got_Latn, hbb_Latn, tll_Latn, bug_Latn, kxp_Arab, qaa_Latn, krr_Khmr, kjg_Laoo, isu_Latn, kmu_Latn, gof_Latn, sdk_Latn, mne_Latn, baw_Latn, idt_Latn, xkg_Latn, mgo_Latn, dtr_Latn, kms_Latn, ffm_Latn, hna_Latn, nxl_Latn, bfd_Latn, odk_Arab, miq_Latn, mhx_Latn, kam_Latn, yao_Latn, pnt_Grek, kby_Latn, kpv_Latn, kbx_Latn, cim_Latn, qvo_Latn, pih_Latn, nog_Latn, nco_Latn, rmy_Cyrl, clo_Latn, dmg_Latn, aaa_Latn, rel_Latn, ben_Latn, loh_Latn, thl_Deva, chd_Latn, cni_Latn, cjs_Latn, lbe_Latn, ybh_Deva, zxx_Zyyy, awa_Latn, gou_Latn, xmm_Latn, nqo_Latn, rut_Cyrl, kbq_Latn, tkr_Latn, dwr_Ethi, ckt_Cyrl, ady_Latn, yea_Mlym, nhx_Latn, niv_Cyrl, bwt_Latn, xmg_Latn, chy_Latn, mfj_Latn, hre_Latn, bbk_Latn, shn_Latn, lrc_Latn, qvc_Latn, muv_Mlym, mdr_Latn, luy_Latn, lzh_Hani, fuh_Latn, mle_Latn, brx_Deva, pex_Latn, kau_Latn, yrk_Latn, hin_Latn, ekm_Latn, msb_Latn, unr_Orya, cac_Latn, chp_Cans, ckt_Latn, bss_Latn, lts_Latn, bbj_Latn, ttt_Cyrl, kwu_Latn, smn_Cyrl, kpy_Cyrl, tod_Latn, wbm_Latn, tcy_Latn, arc_Syrc, nst_Latn, tuz_Latn, bob_Latn, bfn_Latn, pli_Deva, snl_Latn, kwd_Latn, lgg_Latn, nza_Latn, wbr_Deva, lan_Latn, kmz_Latn, bzi_Thai, hao_Latn, nla_Latn, qxr_Latn, ken_Latn, tbj_Latn, blk_Latn, ybb_Latn, nwe_Latn, gan_Hani, snk_Latn, kak_Latn, tpl_Latn, hla_Latn, tks_Arab, pea_Latn, bya_Latn, enc_Latn, jgo_Latn, tnp_Latn, aph_Deva, bgf_Latn, brv_Laoo, nod_Thai, niq_Latn, nwi_Latn, xmd_Latn, gbj_Orya, syr_Latn, ify_Latn, xal_Latn, bra_Deva, cgc_Latn, bhs_Latn, pwg_Latn, ang_Runr, oki_Latn, qve_Latn, qvm_Latn, bkm_Latn, bkh_Latn, niv_Latn, zuh_Latn, mry_Latn, fiu_Cyrl, ssn_Latn, rki_Mymr, sox_Latn, yav_Latn, nyo_Latn, dag_Arab, qxh_Latn, bze_Latn, myx_Latn, zaw_Latn, ddg_Latn, wnk_Latn, bwx_Latn, mqy_Latn, lad_Hebr, boz_Latn, lue_Latn, ded_Latn, pli_Latn, avk_Cyrl, wms_Latn, sgd_Latn, azn_Latn, ajz_Latn, psp_Latn, jra_Latn, smt_Latn, ags_Latn, csw_Latn, wtk_Latn, emp_Cyrl, koi_Latn, tkr_Cyrl, amp_Latn, ymp_Latn, mfh_Latn, tdb_Deva, omw_Latn, khb_Talu, doi_Deva, gld_Cyrl, ava_Latn, chu_Latn, dnw_Latn, azo_Latn, dug_Latn, bce_Latn, kmr_Latn, kpy_Armn, abq_Cyrl, trp_Latn, ewo_Latn, the_Deva, hig_Latn, pkb_Latn, mxu_Latn, oji_Latn, tnt_Latn, mzm_Latn, mns_Cyrl, lbe_Cyrl, qvh_Latn, kmg_Latn, sps_Latn, brb_Khmr, tah_Latn, sxb_Latn, mkz_Latn, mgq_Latn, got_Goth, lns_Latn, arc_Latn, akb_Latn, skr_Latn, nsk_Cans, sml_Latn, pce_Mymr, eee_Thai, lhm_Deva, yux_Cyrl, bqm_Latn, bcc_Arab, nas_Latn, agq_Latn, xog_Latn, tsb_Latn, fub_Latn, mqj_Latn, nsk_Latn, bxr_Latn, dln_Latn, ozm_Latn, rmy_Latn, cre_Cans, kim_Cyrl, cuh_Latn, ngl_Latn, yas_Latn, bud_Latn, miy_Latn, ame_Latn, pnz_Latn, raj_Deva, enb_Latn, cmo_Khmr, saq_Latn, tpu_Khmr, eve_Cyrl, cdo_Hani

Table 7: Languages by resource groups categorized by counting the number of tokens in the MaLA monolingual corpus (Ji et al., 2024a). "Unseen" means those languages are not used for continual pre-training in this paper.

them if the average line length is less than 80, the maximum line length is less than 120, and the alphanumeric fraction is more than 45%.

Code-Adjacent Procedural Data We augment our data mix with procedural code-adjacent data ranging from instructive data such as StackOverflow QnAs, library documentation and tutorials [12] to Jupyter notebooks, synthetic code textbooks (Su et al., 2024), GitHub commits, and issue descriptions (Lozhkov et al., 2024). We further mirror Paul et al. (2025) in sourcing parallel math problem solutions to code data.

[12] https://huggingface.co/datasets/BEE-spoke-data/code-tutorials-en

Type	Category	Sample Rate	Token Counts		Percentage of Mixes	
			Original	Final	Bilingual	Monolingual
instruction	EN	0.1	9,204,199,807	920,419,981	0.32%	0.47%
	high	0.2	39,403,448,029	7,880,689,606	2.72%	4.01%
	medium-high	0.5	30,651,187,534	15,325,593,767	5.28%	7.81%
	medium	5.0	1,444,764,863	7,223,824,315	2.49%	3.68%
	medium-low	20.0	47,691,495	953,829,900	0.33%	0.49%
	low	50.0	3,064,796	153,239,800	0.05%	0.08%
	code/reasoning	1.0	612,208,775	612,208,775	0.21%	0.31%
code	code	1.0	43,478,432,765	43,478,432,765	14.99%	22.15%
book	Gutenberg	1.0	5,173,357,710	5,173,357,710	1.78%	2.64%
paper	EN	1.0	38,256,934,181	38,256,934,181	13.19%	19.49%
	ZH	1.0	61,787,372	61,787,372	0.02%	0.03%
monolingual	EN	0.1	3,002,029,817	300,202,982	0.10%	0.15%
	high	0.5	40,411,201,964	20,205,600,982	6.97%	10.29%
	medium-high	1.0	27,515,227,962	27,515,227,962	9.49%	14.02%
	medium	5.0	2,747,484,380	13,737,421,900	4.74%	7.00%
	medium-low	20.0	481,935,633	9,638,712,660	3.32%	4.91%
	low	50.0	97,535,696	4,876,784,800	1.68%	2.48%
bilingual	very high	0.1	85,001,097,362	4,250,054,868	1.47%	0.00%
	high	0.1	207,688,940,222	20,768,894,022	7.16%	0.00%
	medium-high	0.2	46,777,497,823	9,355,499,565	3.23%	0.00%
	medium	0.5	64,375,100,302	32,187,550,151	11.10%	0.00%
	medium-low	1.0	20,361,578,347	20,361,578,347	7.02%	0.00%
	low	2.0	2,503,240,829	5,006,481,658	1.73%	0.00%
	very low	10.0	175,309,923	1,753,099,230	0.60%	0.00%

Table 8: Composition of training data mixtures showing original and final token counts by data type, with sampling rates and distribution percentages across bilingual and monolingual configurations.

C Detailed Evaluation Setup

C.1 Tasks and Benchmarks

Table 9 summarizes the tasks and benchmarks used for evaluation.

Text Classification SIB-200 (Adelani et al., 2023) and Taxi-1500 (Ma et al., 2023) are two representative multilingual topic-classification benchmarks. SIB-200 is based on Flores-200, covering 205 languages and dialects. The topics are "science/technology", "travel", "politics", "sports", "health", "entertainment", and "geography". Taxi-1500 is a sentence classification data set with 6 topics, i.e., recommendation, faith, description, sin, grace, and violence, collected from the Parallel Bible Corpus (Mayer and Cysouw, 2014). It covers 1,502 typologically diverse languages spanning 112 language families.

Commonsense Reasoning We evaluate `EMMA-500` models' commonsense-reasoning skills with two multilingual benchmarks. XCOPA (Ponti et al., 2020) targets causal inference across 11 languages, XStoryCloze (Lin et al., 2022) gauges narrative completion in 11 languages.

Natural Language Inference We evaluate on XNLI (Conneau et al., 2018), where sentence pairs in different languages need to be classified as entailment, contradiction, or neutral.

Machine Translation FLORES-200 (Costa-jussà et al., 2022) is a multilingual benchmark designed to evaluate machine translation performance across 204 language pairs involving English, yielding 408 translation directions with an emphasis on low-resource settings.

Text Summarization We use two multilingual news summarization datasets: MassiveSumm (Varab and Schluter, 2021) and XL-Sum (Hasan et al., 2021). We subsample MassiveSumm into two sets, i.e., **MassiveSumm long**[13] designed for longer texts, aiming for a median of 5500 tokens within a range of 3500 to 7500 tokens, and **MassiveSumm short**[14] on shorter texts, targeting a median of 1200 tokens with a maximum of 1500 tokens. For practical reasons, we ensure a balanced number of documents across all languages, with a minimum of 100 and a maximum of 2500 documents per language in both MassiveSumm subsets. Additionally, we utilized the XL-Sum, a diverse dataset containing over one million professionally annotated article-summary pairs across 44 languages from the BBC.

Machine Reading Comprehension We use two datasets of machine reading comprehension: BELEBELE (Bandarkar et al., 2023) and the multilingual ARC challenge. The BELEBELE dataset (Bandarkar et al., 2023) is a parallel multilingual machine reading comprehension benchmark span-

[13] https://huggingface.co/datasets/MaLA-LM/MassiveSumm_long

[14] https://huggingface.co/datasets/MaLA-LM/MassiveSumm_short

ning 122 languages across resource levels. Each of its carefully validated multiple-choice questions (with four options per question) derives from FLORES-200 passages, testing nuanced comprehension while maintaining full parallelism for cross-lingual comparison. The AI2 Reasoning Challenge (ARC) dataset (Clark et al., 2018) is a rigorously constructed benchmark for advanced question answering and reasoning evaluation. It comprises 7,787 genuine grade-school level science questions. We use the multilingual ARC translated by Lai et al. (2023) using ChatGPT.

Math We use the Multilingual Grade School Math Benchmark (MGSM) (Shi et al., 2022) that extends the GSM8K dataset (Cobbe et al., 2021) to evaluate cross-lingual mathematical reasoning in large language models. Derived from GSM8K's collection of 8,500 linguistically diverse grade-school math word problems requiring multi-step reasoning, MGSM comprises 250 carefully selected problems manually translated into ten typologically diverse languages.

The evaluation benchmarks are abbreviated as follows:

- `XC`: XCOPA (Cross-lingual Choice of Plausible Alternatives)
- `XSC`: XStoryCloze (Cross-lingual Story Cloze Test)
- `BELE`: BELEBELE, Multilingual Reading Comprehension Benchmark
- `ARC`: Multilingual AI2 Reasoning Challenge

For the MGSM benchmark, we distinguish between:

- `Dir.`: Direct prompting (question-only)
- `CoT`: Chain-of-Thought prompting (with reasoning steps) (Wei et al., 2022)

C.2 Evaluation Software

We evaluate benchmarks using the `lm-evaluation-harness` framework (Gao et al., 2023) for supported tasks, and custom or open-source scripts for others. For text classification (e.g., SIB-200, Taxi-1500), we predict the highest-probability category using next-token probabilities via Transformers (Wolf et al., 2019). For translation and open-ended generation, we use vLLM (Kwon et al., 2023) for faster inference.

C.3 Baselines

We compare our models with several groups of baseline approaches, covering both established and emerging LLMs.

Llama 2 series including CPT models:

- Llama 2 base & chat (Touvron et al., 2023b): the second generation of Meta's foundational 7BB parameter models release in 2023;
- CodeLlama 2 (Roziere et al., 2023): continual pre-trained model specialized for code;
- LlaMAX Llama 2 base and Alpaca (Lu et al., 2024): continual pre-trained and instruction-tuned variants on Alpaca (Taori et al., 2023);
- MaLA-500 Llama 2 v1 & v2 (Lin et al., 2024): multilingual adaptations with continual pre-training and low-rank adaptation (Hu et al., 2022);
- YaYi Llama 2[15]: a Chinese-optimized version;
- TowerBase and TowerInstruct Llama 2 (Alves et al., 2024): domain-specific adaptations optimized mainly for machine translation;
- EMMA-500 Llama 2 (Ji et al., 2024a): a massively multilingual variant.

Multilingual LLMs including recent advances:

- Occiglot Mistral v0.1 base and instruct [16]: multilingual variants continue-trained and instruct-tuned on Mistral (Jiang et al., 2023);
- BLOOM & BLOOMZ (Scao et al., 2022; Muennighoff et al., 2022): early open multilingual and instruction-tuned models released in 2022;
- YaYi [17]: a model with Chinese-focused multilingual capabilities;
- Aya 23 (Aryabumi et al., 2024): a multilingual instruction-tuned LLM supporting 23 languages;
- Aya Expanse (Dang et al., 2024): a multilingual variant optimized with various post-training methods such as synthetic data augmentation, iterative preference training, and model merging;
- Gemma series: Google's multilingual models including generations 1[18] & 2 (Team et al., 2024);
- Qwen series (Bai et al., 2023): Alibaba's multilingual offerings including generations 1.5[19],

[15] http://https://huggingface.co/wenge-research/yayi-7b-llama2

[16] http://https://huggingface.co/occiglot

[17] http://https://huggingface.co/wenge-research/yayi-7b

[18] http://https://huggingface.co/google/gemma-7b

[19] http://https://huggingface.co/Qwen/Qwen1.5-7B

Tasks	Dataset	Metric	Samples/Lang	N Lang	Domain
Text Classification (Appendix D.2)	SIB200 (Adelani et al., 2023)	Accuracy	204	205	Misc
	Taxi1500 (Ma et al., 2023)	Accuracy	111	1507	Bible
Commonsense Reasoning (Appendix D.3)	XCOPA (Ponti et al., 2020)	Accuracy	600	11	Misc
	XStoryCloze (Lin et al., 2022)	Accuracy	1870	11	Misc
Natural Language Inference (Appendix D.4)	XNLI (Conneau et al., 2018)	Accuracy	2490	15	Misc
Machine Translation (Appendix D.5)	FLORES-200 (Costa-jussà et al., 2022)	BLEU, chrF++	1012	204	Misc
Summarization (Appendix D.6)	XL-Sum (Hasan et al., 2021)	ROUGE-L, BERTScore	2537	44	News
	MassiveSumm Long (Varab and Schluter, 2021)	ROUGE-L, BERTScore	3908	55	News
	MassiveSumm Short (Varab and Schluter, 2021)	ROUGE-L, BERTScore	5538	88	News
Machine Comprehension (Appendix D.7)	BELEBELE (Bandarkar et al., 2023)	Accuracy	900	122	Misc
	ARC multilingual (Lai et al., 2023)	Accuracy	1170	31	Misc
Math (Appendix D.8)	MGSM direct (Shi et al., 2022)	Accuracy	250	10	Misc
	MGSM CoT (Shi et al., 2022)	Accuracy	250	10	Misc

Table 9: Evaluation statistics. Sample/Lang: average number of test samples per language; N Lang: number of languages covered.

2 (Yang et al., 2024), & 2.5 (Qwen et al., 2025);
- Macro-LLM (Ming et al., 2024): a continue-pretrained LLM using Qwen2.

Llama 3 series including CPT models:
- Llama 3 (Dubey et al., 2024): Meta's latest generation with a cutoff or March 2023;
- Llama 3.1 (Dubey et al., 2024): Refined version with improved capabilities, which extends the data cutoff to December 2023;
- LLaMAX Llama 3 base and Alpaca (Lu et al., 2024): continue-trained and instruction-tuned variants.

Our evaluation encompasses these representative models to ensure a comprehensive comparison. The selection covers diverse architectures and linguistic capabilities. We continue to expand our benchmarking coverage and welcome collaboration inquiries regarding additional model evaluations.

D Detailed Results

This section presents the detailed results of each evaluated task, including average scores categorized by language resource groups and per-language scores if the benchmark has a few languages. For massively multilingual benchmarks with more than a dozen of languages, we provide the results on our project website (https://mala-lm.github.io/emma-500-gen2). We also provide all the model outputs of different generation tasks on the project website.

D.1 Detailed Overall Results

Tables 10 and 11 show the detailed overall performance of deterministic and generation tasks, averaged across all languages evaluated.

D.2 Text Classification

We evaluate with 3-shot prompting on the SIB-200 and Taxi-1500 text classification datasets. The prompt template for SIB-200 is as follows:

```
Topic Classification: science/
    technology, travel, politics,
    sports, health, entertainment,
    geography.
{examples}
The topic of the news "${text}" is
```

For Taxi-1500, the prompt template is as follows:

```
Topic Classification:
    Recommendation, Faith,
    Description, Sin, Grace,
    Violence.
{examples}
The topic of the verse "${text}" is
```

As for SIB-200 task, Table 12 shows that `EMMA-500 Llama 3.1 8B Bi` attains the highest accuracy among `EMMA-500` model series (62.07%), edging its backbone by +0.65% and crossing the 60% threshold together with `EMMA-500 Llama 3 8B Mono` (60.62%). Yet the other two variants fall significantly below their bases: `EMMA-500` Llama 3 8B Bi drops to 39.40%, while `EMMA-500` Llama 3.1 8B Mono slips to 26.16%. The mixed picture—improvements for some settings, regressions for others—suggests that our current continual pre-training strategy for `EMMA-500` models may not be consistently helpful for SIB-200 task, and calls for deeper analysis in terms of data mix and training strategy.

With its religious text domain, short inputs and 1500+ languages, Taxi-1500 is markedly harder.

Model	Classification		Commonsense		NLI	Comprehension		Math	
	SIB	Taxi	XC	XSC	XNLI	BELE	ARC	Dir.	CoT
Llama 2 7B	22.41	17.54	56.67	57.55	40.19	26.27	27.56	6.69	6.36
Llama 2 7B Chat	25.58	15.44	55.85	58.41	38.58	**29.05**	28.02	10.22	10.91
CodeLlama 2 7B	23.35	17.03	54.69	55.68	40.19	27.38	25.23	5.93	6.64
LLaMAX Llama 2 7B	10.61	23.52	54.38	60.36	44.27	23.09	26.09	3.35	3.62
LLaMAX Llama 2 7B Alpaca	27.89	15.09	56.60	63.83	45.09	24.48	31.06	5.05	6.35
MaLA-500 Llama 2 10B v1	23.25	**25.27**	53.09	53.07	38.11	22.96	21.16	0.91	0.73
MaLA-500 Llama 2 10B v2	19.30	23.39	53.09	53.07	38.11	22.96	21.16	0.91	0.73
YaYi Llama 2 7B	24.57	17.73	56.71	58.42	41.28	28.32	28.40	7.09	7.22
TowerBase Llama 2 7B	19.34	17.73	56.33	57.78	39.84	26.36	27.94	6.15	6.16
TowerInstruct Llama 2 7B	20.53	17.29	57.05	59.24	40.36	27.93	**30.10**	7.24	8.24
EMMA-500 Llama 2 7B	**31.27**	19.82	**63.11**	**66.38**	**45.14**	26.75	29.53	**17.02**	**18.09**
Occiglot Mistral 7B v0.1	32.69	22.26	56.67	58.10	42.35	30.16	29.77	13.31	14.07
Occiglot Mistral 7B v0.1 Instruct	34.31	18.76	56.55	59.39	40.81	32.05	30.88	22.76	22.16
BLOOM 7B	17.82	14.76	56.89	59.30	41.60	24.11	23.65	2.87	2.29
BLOOMZ 7B	29.73	16.96	54.87	57.12	37.13	39.32	23.95	2.55	2.15
YaYi 7B	35.76	16.12	56.64	60.67	39.87	37.97	24.44	2.76	3.02
Aya 23 8B	41.50	**22.64**	55.13	60.93	43.12	40.08	31.08	22.29	24.71
Aya Expanse 8B	57.01	18.73	56.38	64.80	**45.56**	46.98	36.56	43.02	41.45
Gemma 7B	58.21	13.83	63.64	65.01	42.58	43.37	38.68	38.22	35.78
Gemma 2 9B	46.25	18.05	66.33	**67.67**	46.74	**54.49**	**44.15**	32.95	44.69
Qwen 1.5 7B	47.95	7.29	59.44	59.85	39.47	41.83	28.93	31.56	30.36
Qwen 2 7B	54.95	21.87	60.31	61.46	42.77	49.31	33.82	48.95	51.47
Qwen 2.5 7B	53.89	17.87	61.71	62.06	43.31	54.11	35.30	**53.78**	**55.60**
Marco-LLM GLO 7B	**64.15**	21.99	**62.45**	63.87	43.99	53.95	36.34	51.85	52.02
Llama 3 8B	**63.70**	**21.73**	61.71	63.41	44.97	40.73	34.80	27.45	**28.13**
Llama 3.1 8B	61.42	20.20	61.71	63.58	45.62	**45.19**	**34.93**	**28.36**	27.31
LLaMAX Llama 3 8B	48.60	23.01	63.04	64.31	44.13	36.96	33.54	20.80	19.96
LLaMAX Llama 3 8B Alpaca	58.97	17.71	**64.36**	**68.27**	45.08	39.41	34.53	14.18	17.16
EMMA-500 Llama 3 8B Mono	60.62	22.32	66.20	67.36	44.15	39.73	33.22	23.53	25.33
EMMA-500 Llama 3 8B Bi	39.40	25.13	66.82	68.35	**45.15**	**39.84**	**34.84**	23.49	26.29
EMMA-500 Llama 3.1 8B Mono	26.16	19.71	65.38	67.64	39.98	38.86	34.00	**24.95**	**27.35**
EMMA-500 Llama 3.1 8B Bi	62.07	24.87	**67.25**	**68.47**	44.67	37.00	34.59	23.85	25.76

Table 10: Overall results of deterministic tasks including text classification, commonsense reasoning, natural language inference, reading comprehension, and math reasoning. XC: XCOPA; XSC: XStoryCloze; BELE: BELEBELE; ARC: the multilingual AI2 Reasoning Challenge; Dir.: MGSM by direct prompting; CoT: MGSM by CoT prompting. **Underline and bold** represents the absolute best, underline means the second best, and **bold** signifies the best within a specific group. Our EMMA-500 models are among the top 2 models on text classification and commonsense reasoning.

Here, the advantages of EMMA-500 are more pronounced. Where most 8 B models struggle to reach 20% accuracy, all four EMMA-500 variants hover around or above that mark. EMMA-500 Llama-3 8B-Bi achieves 25.13%, a gain of 3.40% over the base model, and EMMA-500 Llama-3.1 8B-Bi reaches 24.87%.

D.3 Commonsense Reasoning

In the commonsense reasoning evaluation, all evaluations are zero-shot; accuracy is the evaluation metric. Languages are bucketed into groups according to language availability and possible corpus size, following Ji et al. (2024a).

Across XCOPA and XStoryCloze evaluation results (Table 13), the four EMMA-500 variants built on Llama-3/3.1 8 B clearly lift commonsense-reasoning accuracy over their backbones. The best model, EMMA-500 Llama 3.1 8B Bi, averages 67.25% on XCOPA and 68.47% on XStoryCloze—roughly +5% over vanilla Llama-3.1 and almost ahead of, all other similar size models. The performance gains are consistent across all different languages (Tables 14 and 15). Besides, improvements concentrate in medium- and low-resource languages: EMMA-500 lifts Llama-3/3.1 by roughly +7% in medium-resource languages and flattens the usual low-resource performance drop on XStoryCloze (64% vs. 53% for the base), while high-resource accuracy remains competitive at around 70%.

One more interesting finding is that EMMA-500 monolingual variants are slightly behind bilingual variants by about 1%, confirming that balanced bilingual continual training adds marginal yet consistent performance gains.

D.4 Natural Language Inference

According to the XNLI evaluation results shown in Table 16, the bilingual EMMA-500 8B variant outperform its backbone on average (45.15% vs 44.97%) and lift low-resource accuracy by roughly +4%, closing most of the gap to the stronger Gemma 2 9 B while matching Aya Expanse 8 B. The monolingual Llama-3 variant keeps pace (44.15%), but the same Llama backbone models with monolin-

Model	Flores200				MassiveSumm-L		MassiveSumm-S		XL-Sum	
	chrF++	BLEU	chrF++	BLEU	R-L	BERT	R-L	BERT	R-L	BERT
Llama 2 7B	15.13	4.62	30.32	12.93	4.74	63.89	7.85	65.35	7.11	66.52
Llama 2 7B Chat	16.95	4.95	31.72	12.28	4.73	63.52	9.76	67.01	8.84	68.44
CodeLlama 2 7B	14.94	4.27	28.57	10.82	5.63	64.51	7.59	64.83	7.15	65.74
LLaMAX Llama 2 7B	7.42	0.80	13.66	1.99	4.56	62.69	5.22	63.06	5.29	64.59
LLaMAX Llama 2 7B Alpaca	28.35	12.51	42.27	22.29	4.61	62.76	10.71	**67.92**	**10.11**	**69.24**
MaLA-500 Llama 2 10B v1	6.08	0.60	13.60	2.29	4.39	64.50	4.97	63.51	5.45	63.96
MaLA-500 Llama 2 10B v2	6.38	0.54	15.44	2.87	4.37	**64.66**	5.02	63.75	5.44	64.28
YaYi Llama 2 7B	14.87	4.41	31.38	12.98	**4.98**	64.17	7.80	65.24	7.98	67.21
TowerBase Llama 2 7B	16.03	4.83	31.47	13.74	4.81	64.51	8.11	65.53	7.65	67.09
TowerInstruct Llama 2 7B	15.64	3.23	25.43	4.81	4.82	64.61	10.14	67.76	8.89	68.46
EMMA-500 Llama 2 7B	**33.25**	**15.58**	**45.78**	**25.37**	4.79	63.80	8.32	65.14	8.58	67.20
Occiglot Mistral 7B v0.1	16.10	4.32	31.13	13.12	5.14	63.95	8.16	63.65	7.33	66.20
Occiglot Mistral 7B v0.1 Instruct	15.80	3.99	31.65	11.61	5.16	63.50	7.82	63.79	8.31	66.96
BLOOM 7B	11.80	2.81	27.84	9.57	4.88	64.36	6.79	62.30	6.99	64.78
BLOOMZ 7B	16.10	7.44	34.74	20.22	2.91	57.20	3.28	29.75	11.15	69.82
YaYi 7B	13.50	4.37	21.36	4.82	4.95	64.24	8.28	65.44	**12.06**	69.74
Aya 23 8B	16.15	6.46	32.36	13.87	6.33	65.94	8.43	65.85	8.68	66.79
Aya Expanse 8B	23.89	6.88	36.86	13.12	7.44	**67.66**	**9.24**	67.68	10.51	68.73
Gemma 7B	23.05	9.05	43.68	23.79	6.18	62.14	8.35	62.25	6.70	64.52
Gemma 2 9B	**26.48**	**12.09**	38.87	23.15	5.86	59.70	7.86	58.11	7.38	65.45
Qwen 1.5 7B	17.77	5.87	35.87	15.58	6.09	59.19	8.49	62.70	9.58	69.13
Qwen 2 7B	17.17	5.56	37.61	17.39	6.65	56.31	8.63	56.14	10.18	69.35
Qwen 2.5 7B	17.49	5.72	38.89	18.95	7.44	61.62	9.04	58.91	10.41	69.69
Marco-LLM GLO 7B	23.34	9.27	**44.57**	**25.17**	6.57	57.15	8.10	57.45	11.46	**70.41**
Llama 3 8B	24.08	9.93	43.72	23.78	5.10	55.58	6.44	50.98	8.47	67.08
Llama 3.1 8B	24.69	10.11	44.10	24.19	5.41	56.09	6.77	54.96	8.57	66.97
LLaMAX Llama 3 8B	4.65	0.45	4.66	0.48	6.00	65.90	8.77	66.46	8.28	66.72
LLaMAX Llama 3 8B Alpaca	**26.86**	**11.64**	**45.45**	**25.10**	**7.62**	67.64	**12.44**	**68.95**	**11.39**	69.99
EMMA-500 Llama 3 8B Mono	38.34	20.33	50.86	30.38	5.38	61.06	7.18	**63.46**	9.11	66.12
EMMA-500 Llama 3 8B Bi	**42.15**	**24.02**	**54.33**	**34.40**	**5.57**	59.23	**7.74**	63.18	9.10	66.72
EMMA-500 Llama 3.1 8B Mono	37.41	19.44	48.12	27.57	5.44	**61.89**	7.21	63.36	**9.66**	**67.21**
EMMA-500 Llama 3.1 8B Bi	42.07	23.86	53.49	33.64	4.78	58.47	6.67	61.64	8.56	65.90

Table 11: Overall results of generation tasks including text summarization and machine translation. BERT: BERTScore; R-L: ROUGE-L. **Underline and bold** represents the absolute best, underline means the second best, and **bold** signifies the best within a specific group. Our EMMA-500 models trained with bilingual mix are the best-performing models on machine translation.

gual continual training leads to a performance drop (39.98% in EMMA-500 Llama 3.1 8B Mono), leaving the bilingual Llama-3.1 model (44.67%) as the only Llama-3.1-based EMMA-500 choice that preserves gains in low-resource languages without a drop in high-resource languages. Overall, EMMA-500 offers modest NLI benefits concentrated in low-resource languages.

D.5 Machine Translation

We evaluate all models for machine translation under a 3-shot prompting setup. Specifically, we use the prompt below.

```
Translate the following sentence
    from {src_lang} to {tgt_lang}
[{src_lang}]: {src_sent}
[{tgt_lang}]:
```

Translation quality is assessed using BLEU (Papineni et al., 2002) and chrF++ (Popović, 2015), implemented via sacrebleu(Post, 2018). BLEU is computed using the flores200 tokenizer, which allows consistent segmentation even for languages lacking explicit word boundaries; chrF++ is calculated with a word n-gram order of 2. For transparency and reproducibility, we report full metric signatures. [20]

On translations from all other languages to English (Table 18), EMMA-500 variants based on both Llama-3 8B and Llama-3.1 8B show clear, systematic gains under monolingual and bilingual continual training. Specifically, our continual training strategy boosts the BLEU/chrF++ scores of Llama-3 8B from 23.78/43.72 to 30.38/50.86 (EMMA-500 Llama 3 8B Mono, +6.6 BLEU) and further to 34.4/54.33 (EMMA-500 Llama 3 8B Bi, +10.6 BLEU); the same schedule lifts Llama-3.1 8B from 24.19/44.1 to 27.57/48.12 (EMMA-500 Llama 3.1 8B Mono, +3.4 BLEU) and 33.6/53.5 ((EMMA-500 Llama 3 8B Bi, +9.5 BLEU).

Gains are even larger when English is the source language (Table 19): EMMA-500 Llama 3 Bi more than doubles base-model BLEU (9.93 → 24.02) and ChrF++ (24.08 → 42.15), while EMMA-500 Llama 3.1 Bi shows a comparable jump (10.11 → 23.86 BLEU). Improvements are greatest in the medium-low and low-resource languages, confirming that language-balanced bilingual contin-

[20] BLEU: nrefs:1|case:mixed|eff:no|tok:flores200|smooth:exp |version:2.4.2; chrF++: nrefs:1|case:mixed|eff:yes|nc:6|nw:2| space:no|version:2.4.2

Model	SIB-200						Taxi-1500					
	Avg	**High**	**Med-High**	**Medium**	**Med-Low**	**Low**	**Avg**	**High**	**Med-High**	**Medium**	**Med-Low**	**Low**
Llama 2 7B	22.41	25.98	22.17	22.43	23.36	22.74	17.54	19.57	17.78	17.89	18.17	18.04
Llama 2 7B Chat	25.58	29.15	25.27	25.43	26.58	26.02	15.44	18.78	15.67	15.77	16.09	16.01
CodeLlama 7B	23.35	25.80	23.01	23.39	24.27	23.84	17.03	17.47	17.11	17.12	17.17	17.15
LLaMAX Llama 2 7B	10.61	12.83	10.96	10.66	11.11	10.78	23.52	23.08	23.54	23.55	23.52	23.53
LLaMAX Llama 2 7B Alpaca	27.89	33.36	27.69	27.88	29.74	28.66	15.09	18.87	15.31	15.43	15.70	15.57
MaLA-500 10B V1	23.25	23.00	23.44	23.03	23.30	23.13	25.27	23.53	25.08	25.00	24.83	24.76
MaLA-500 10B V2	19.30	17.83	18.57	18.42	19.51	19.04	23.39	21.27	23.35	23.07	22.94	22.72
Yayi Llama 2 7B	24.57	27.99	24.08	24.64	25.72	25.03	17.73	18.62	17.84	17.90	18.01	17.98
Tower Base Llama 2 7B V0.1	19.34	21.30	18.90	19.07	19.86	19.53	17.73	18.74	17.76	17.87	18.04	18.06
Tower Instruct Llama 2 7B V0.2	20.53	22.91	20.22	20.46	21.10	20.81	17.29	20.16	17.69	17.72	18.05	17.92
EMMA-500 Llama 2 7B	31.27	33.02	30.34	30.86	31.93	31.97	19.82	24.32	21.07	21.32	21.68	21.73
Occiglot Mistral 7B	32.69	37.77	32.91	32.87	34.07	33.27	22.26	24.74	22.53	22.61	22.69	22.62
Occiglot Mistral 7B Instruct	34.31	37.79	34.01	34.20	35.37	34.87	18.76	24.59	19.41	19.42	19.68	19.59
BLOOM 7B1	17.82	21.57	19.54	18.61	18.13	17.73	14.76	15.71	14.98	14.93	14.91	14.93
BLOOMZ 7B1	29.73	29.77	29.95	30.13	29.89	29.99	16.96	16.89	16.99	16.98	17.00	17.00
Yayi 7B	35.76	39.51	36.50	36.62	36.40	36.40	16.12	16.56	16.17	16.24	16.26	16.19
Aya 23 8B	41.50	45.45	42.17	42.43	43.42	42.20	22.64	22.82	22.63	22.56	22.56	22.52
Aya Expanse 8B	57.01	63.82	59.73	58.99	59.84	58.06	18.73	19.56	19.07	19.08	19.03	19.08
Gemma 7B	58.21	68.76	60.00	60.14	62.08	59.97	13.83	25.10	14.78	15.42	16.92	16.55
Gemma 2 9B	46.25	51.94	45.79	46.78	47.82	47.19	18.05	29.88	19.43	20.27	21.72	21.48
Qwen 1.5 7B	47.95	55.88	49.42	49.53	50.34	48.94	7.29	12.42	7.61	7.75	8.39	8.18
Qwen 2 7B	54.95	66.84	58.04	57.27	58.33	56.06	21.87	27.85	22.42	22.66	23.16	23.01
Qwen 2.5 7B	53.89	64.57	56.45	56.08	56.82	54.85	17.87	18.12	18.01	17.95	17.85	17.82
Macro-LLM GLO 7B	64.15	72.73	65.98	65.73	66.96	65.07	21.99	21.31	21.84	21.92	21.93	21.97
Llama 3 8B	63.70	73.89	66.11	66.33	67.89	65.53	21.73	32.43	22.64	23.07	23.79	23.61
Llama 3.1 8B	61.42	70.83	63.42	63.67	65.33	63.17	20.20	28.22	21.12	21.59	22.22	22.02
LLaMAX Llama 3 8B	48.60	51.41	48.19	48.94	50.16	49.70	23.01	22.32	22.84	22.85	22.82	22.91
LLaMAX Llama 3 8B Alpaca	58.97	70.50	60.57	60.96	63.71	60.91	17.71	30.56	18.56	19.32	20.83	20.48
EMMA-500 Llama 3 8B Mono	60.62	60.72	58.70	60.26	62.10	62.34	22.32	23.13	22.04	22.02	22.10	22.16
EMMA-500 Llama 3 8B Bi	39.40	37.14	39.83	38.88	38.39	39.65	25.13	29.53	26.13	26.45	26.71	26.72
EMMA-500 Llama 3.1 8B Mono	26.16	25.31	25.74	25.84	25.80	26.17	19.71	18.82	19.41	19.45	19.47	19.62
EMMA-500 Llama 3.1 8B Bi	62.07	59.80	60.15	61.60	63.31	63.33	24.87	25.71	25.50	25.45	25.40	25.42

Table 12: 3-shot results on SIB-200 and Taxi-1500 (Accuracy %).

Model	XCOPA				XStoryCloze			
	Avg	High	Medium	Low	Avg	High	Medium	Low
Llama 2 7B	56.67	62.10	53.90	51.60	57.55	63.38	54.45	49.21
Llama 2 7B Chat	55.85	61.25	53.13	50.60	58.41	64.80	54.77	49.74
CodeLlama 2 7B	54.69	58.70	52.53	51.60	55.68	60.68	52.33	49.90
LLaMAX Llama 2 7B	54.38	55.50	54.13	51.40	60.36	64.34	58.82	53.47
LLaMAX Llama 2 7B Alpaca	56.60	59.80	55.17	52.40	63.83	69.08	62.01	54.33
MaLA-500 Llama 2 10B v1	53.09	53.55	53.27	50.20	53.07	58.15	49.22	48.08
MaLA-500 Llama 2 10B v2	53.09	53.55	53.27	50.20	53.07	58.15	49.22	48.08
YaYi Llama 2 7B	56.71	62.10	54.13	50.60	58.42	64.98	54.91	49.04
TowerBase Llama 2 7B	56.33	62.50	52.90	52.20	57.78	64.35	53.67	49.57
TowerInstruct Llama 2 7B	57.05	62.90	54.00	52.00	59.24	66.83	54.53	49.70
EMMA-500 Llama 2 7B	63.11	66.60	62.57	52.40	66.38	68.92	65.73	61.32
Occiglot Mistral 7B v0.1	56.67	62.80	53.37	52.00	58.10	65.18	53.28	50.03
Occiglot Mistral 7B v0.1 Instruct	56.55	62.85	52.97	52.80	59.39	66.94	54.33	50.63
BLOOM 7B	56.89	59.95	55.87	50.80	59.30	61.99	59.05	53.08
BLOOMZ 7B	54.87	56.35	54.60	50.60	57.12	61.14	55.82	49.67
YaYi 7B	56.64	59.55	55.50	51.80	60.67	64.90	59.40	52.65
Aya 23 8B	55.13	59.05	53.30	50.40	60.93	67.27	58.82	49.34
Aya Expanse 8B	56.38	61.50	53.77	51.60	64.80	73.10	61.93	49.80
Gemma 7B	63.64	70.35	61.43	50.00	65.01	69.46	64.49	54.93
Gemma 2 9B	66.33	73.40	64.27	50.40	67.67	72.47	66.69	57.64
Qwen 1.5 7B	59.44	66.85	55.90	51.00	59.85	66.62	56.04	50.56
Qwen 2 7B	60.31	68.65	56.40	50.40	61.46	69.45	56.97	50.50
Qwen 2.5 7B	61.71	72.30	56.63	49.80	62.06	69.41	58.36	51.09
Marco-LLM GLO 7B	62.45	72.20	57.77	51.60	63.87	71.26	61.23	50.66
Llama 3 8B	61.71	68.35	59.03	51.20	63.41	68.50	62.03	53.44
Llama 3.1 8B	61.71	69.30	58.80	48.80	63.58	68.66	62.09	53.87
LLaMAX Llama 3 8B	63.04	67.25	62.30	50.60	64.31	67.78	63.48	57.28
LLaMAX Llama 3 8B Alpaca	64.36	69.55	62.97	52.00	68.27	72.92	67.27	58.64
EMMA-500 Llama 3 8B Mono	66.20	69.25	66.20	54.00	67.36	69.25	66.83	63.70
EMMA-500 Llama 3 8B Bi	66.82	70.80	66.67	51.80	68.35	70.22	67.79	64.82
EMMA-500 Llama 3.1 8B Mono	65.38	68.85	65.13	53.00	67.64	69.64	67.19	63.57
EMMA-500 Llama 3.1 8B Bi	67.25	71.75	66.83	51.80	68.47	70.34	68.13	64.49

Table 13: 0-shot results (Accuracy %) on commonsense reasoning: XCOPA and XStoryCloze.

ual training is especially effective for non-English translations in low-resource settings.

D.6 Text Summarization

We conduct zero-shot text summarization evaluations on three benchmark datasets: MassiveSumm

Model	Avg	et-acc	stderr	ht-acc	stderr	id-acc	stderr	it-acc	stderr	qu-acc	stderr	sw-acc	stderr	ta-acc	stderr	th-acc	stderr	tr-acc	stderr	vi-acc	stderr	zh-acc	stderr
Llama 2 7B	56.67	48.60	2.24	50.60	2.24	62.40	2.17	65.80	2.12	51.60	2.24	52.20	2.24	53.40	2.23	56.20	2.22	54.80	2.23	62.80	2.16	65.00	2.14
Llama 2 7B Chat	55.85	47.80	2.24	50.80	2.24	62.40	2.17	67.20	2.10	50.60	2.24	52.20	2.24	50.60	2.24	55.00	2.23	55.20	2.23	61.20	2.18	61.40	2.18
CodeLlama 2 7B	54.69	46.80	2.23	51.80	2.24	57.40	2.21	63.00	2.16	51.60	2.24	48.80	2.24	55.00	2.23	55.40	2.23	53.80	2.23	55.80	2.22	62.20	2.17
LLaMAX Llama 2 7B	54.38	49.20	2.24	52.60	2.24	53.80	2.23	52.60	2.24	51.40	2.24	54.00	2.23	58.00	2.21	57.20	2.21	53.00	2.23	53.00	2.23	63.40	2.16
LLaMAX Llama 2 7B Alpaca	56.60	51.20	2.24	54.20	2.23	57.20	2.21	61.00	2.18	52.40	2.24	55.00	2.23	57.00	2.22	56.40	2.22	55.20	2.23	55.20	2.23	67.80	2.09
MaLA-500 Llama 2 10B v1	53.09	48.60	2.24	53.40	2.23	53.00	2.23	59.40	2.20	50.20	2.24	52.80	2.23	57.60	2.21	54.20	2.23	51.60	2.24	52.40	2.24	50.80	2.24
MaLA-500 Llama 2 10B v2	53.09	48.60	2.24	53.40	2.23	53.00	2.23	59.40	2.20	50.20	2.24	52.80	2.23	57.60	2.21	54.20	2.23	51.60	2.24	52.40	2.24	50.80	2.24
YaYi Llama 2 7B	56.71	48.80	2.24	50.80	2.24	62.60	2.17	67.00	2.10	50.60	2.24	53.20	2.23	55.20	2.23	54.20	2.23	55.40	2.23	63.20	2.16	62.80	2.16
TowerBase Llama 2 7B	56.33	46.00	2.23	50.20	2.24	60.20	2.19	70.80	2.04	52.20	2.24	50.60	2.24	54.40	2.23	56.00	2.22	53.80	2.23	59.20	2.20	66.20	2.12
TowerInstruct Llama 2 7B	57.05	48.80	2.24	51.60	2.24	62.00	2.17	71.00	2.03	52.00	2.24	51.00	2.24	54.20	2.23	56.40	2.22	54.60	2.23	58.60	2.20	67.40	2.10
EMMA-500 Llama 2 7B	63.11	61.40	2.18	58.00	2.21	74.20	1.96	69.40	2.06	52.40	2.24	66.20	2.12	60.00	2.19	55.60	2.22	62.00	2.17	70.20	2.05	64.80	2.14
Occiglot Mistral 7B v0.1	56.67	47.20	2.23	51.40	2.24	57.00	2.22	74.60	1.95	52.00	2.24	51.60	2.24	57.20	2.21	55.80	2.22	54.40	2.23	55.20	2.23	67.00	2.10
Occiglot Mistral 7B v0.1 Instruct	56.55	46.80	2.23	51.00	2.24	58.40	2.21	73.80	1.97	52.80	2.23	50.00	2.24	56.60	2.22	55.00	2.23	56.20	2.22	55.80	2.22	65.60	2.13
BLOOM 7B	56.89	48.20	2.24	50.80	2.24	69.80	2.06	52.80	2.23	50.80	2.24	51.80	2.24	59.20	2.20	55.40	2.23	51.00	2.24	70.80	2.04	65.20	2.13
BLOOMZ 7B	54.87	49.20	2.24	54.00	2.23	60.60	2.19	51.40	2.24	50.60	2.24	53.40	2.23	57.40	2.21	53.00	2.23	52.20	2.24	59.80	2.19	62.00	2.17
YaYi 7B	56.64	50.40	2.24	53.00	2.23	63.40	2.16	51.80	2.24	51.80	2.24	55.40	2.23	56.20	2.22	54.60	2.23	52.00	2.24	66.40	2.11	68.00	2.09
Aya 23 8B	55.13	50.20	2.24	52.20	2.24	57.00	2.22	60.40	2.19	50.40	2.24	52.20	2.24	55.60	2.22	52.60	2.24	59.60	2.20	58.80	2.20	57.40	2.21
Aya Expanse 8B	56.38	50.80	2.24	51.60	2.24	61.20	2.18	63.40	2.16	51.60	2.24	53.60	2.23	55.20	2.23	50.20	2.24	57.80	2.21	60.20	2.19	64.60	2.14
Gemma 7B	63.64	59.20	2.20	54.80	2.23	72.00	2.01	72.80	1.99	50.00	2.24	60.60	2.19	61.60	2.18	60.40	2.19	65.60	2.13	74.20	1.96	68.80	2.07
Gemma 2 9B	66.33	63.80	2.15	52.80	2.23	77.80	1.86	75.80	1.92	50.40	2.24	63.80	2.15	63.60	2.15	63.80	2.15	67.40	2.10	76.60	1.90	73.80	1.97
Qwen 1.5 7B	59.44	52.00	2.24	53.00	2.23	64.40	2.14	65.20	2.13	51.00	2.24	52.60	2.24	55.80	2.22	57.60	2.21	58.80	2.20	69.20	2.07	74.20	1.96
Qwen 2 7B	60.31	50.80	2.24	50.80	2.24	70.60	2.04	71.40	2.02	50.40	2.24	52.40	2.24	52.80	2.23	61.00	2.18	57.20	2.21	69.40	2.06	76.60	1.90
Qwen 2.5 7B	61.71	49.40	2.24	52.80	2.23	72.60	2.00	74.60	1.95	49.80	2.24	52.20	2.24	54.60	2.23	58.20	2.21	60.00	2.19	75.20	1.93	79.40	1.81
Marco-LLM GLO 7B	62.45	49.60	2.24	52.40	2.24	74.20	1.96	73.40	1.98	51.60	2.24	54.40	2.23	57.20	2.21	58.80	2.20	66.40	2.11	73.80	1.97	75.20	1.93
Llama 3 8B	61.71	53.40	2.23	52.80	2.23	71.40	2.02	71.60	2.02	51.20	2.24	58.00	2.21	60.00	2.19	58.60	2.20	62.40	2.17	71.60	2.02	67.80	2.09
Llama 3.1 8B	61.71	53.00	2.23	53.80	2.23	71.60	2.02	72.60	2.00	48.80	2.24	55.40	2.23	61.20	2.18	57.80	2.21	61.80	2.18	72.40	2.00	70.40	2.04
LLaMAX Llama 3 8B	63.04	61.80	2.18	56.40	2.22	73.00	1.99	68.40	2.08	50.60	2.24	62.60	2.17	60.80	2.19	59.20	2.20	61.40	2.18	70.40	2.04	68.80	2.07
LLaMAX Llama 3 8B Alpaca	64.36	63.00	2.16	54.00	2.23	73.00	1.99	71.80	2.01	52.00	2.24	63.40	2.16	63.20	2.16	61.20	2.18	64.00	2.15	71.80	2.01	70.60	2.04
EMMA-500 Llama 3 8B Mono	66.20	70.80	2.04	61.80	2.18	75.00	1.94	73.40	1.98	54.00	2.23	67.60	2.10	61.80	2.18	60.20	2.19	64.00	2.15	73.00	1.99	66.60	2.11
EMMA-500 Llama 3 8B Bi	66.82	71.40	2.02	60.80	2.19	78.00	1.85	73.40	1.98	51.80	2.24	68.20	2.08	63.40	2.16	58.20	2.21	66.60	2.11	73.20	1.98	70.00	2.05
EMMA-500 Llama 3.1 8B Mono	65.38	65.20	2.13	59.00	2.20	76.00	1.91	73.00	1.99	53.00	2.23	68.40	2.08	62.40	2.17	59.80	2.19	62.80	2.16	74.60	1.95	65.00	2.14
EMMA-500 Llama 3.1 8B Bi	67.25	70.40	2.04	61.00	2.18	77.00	1.88	74.60	1.95	51.80	2.24	69.60	2.06	63.20	2.16	59.80	2.19	67.20	2.10	74.60	1.95	70.60	2.04

Table 14: 0-shot results (ACC %) on XCOPA in all languages

Model	Avg	ar-acc	stderr	en-acc	stderr	es-acc	stderr	eu-acc	stderr	hi-acc	stderr	id-acc	stderr	my-acc	stderr	ru-acc	stderr	sw-acc	stderr	te-acc	stderr	zh-acc	stderr
Llama 2 7B	57.55	49.90	1.29	77.04	1.08	67.37	1.21	50.36	1.29	53.74	1.28	59.23	1.26	48.05	1.29	63.00	1.24	50.50	1.29	54.33	1.28	59.56	1.26
Llama 2 7B Chat	58.41	50.50	1.29	78.69	1.05	67.11	1.21	50.83	1.29	54.07	1.28	59.63	1.26	48.64	1.29	65.52	1.22	52.02	1.29	53.34	1.28	62.21	1.25
CodeLlama 2 7B	55.68	50.10	1.29	71.48	1.16	63.40	1.24	50.43	1.29	49.70	1.29	55.86	1.28	49.37	1.29	59.23	1.26	50.03	1.29	53.74	1.28	59.17	1.26
LLaMAX Llama 2 7B	60.36	58.90	1.27	75.51	1.11	65.25	1.23	54.47	1.28	58.17	1.27	60.62	1.26	52.48	1.29	61.22	1.25	57.18	1.27	59.30	1.26	60.82	1.26
LLaMAX Llama 2 7B Alpaca	63.83	60.36	1.26	81.47	1.00	70.68	1.17	54.86	1.28	62.14	1.25	66.45	1.22	53.81	1.28	67.44	1.21	60.16	1.26	59.30	1.26	65.45	1.22
MaLA-500 Llama 2 10B v1	53.07	48.18	1.29	73.53	1.14	62.41	1.25	49.90	1.29	47.65	1.29	47.92	1.29	46.26	1.28	54.93	1.28	48.71	1.29	52.61	1.28	51.69	1.29
MaLA-500 Llama 2 10B v2	53.07	48.18	1.29	73.53	1.14	62.41	1.25	49.90	1.29	47.65	1.29	47.92	1.29	46.26	1.28	54.93	1.28	48.71	1.29	52.61	1.28	51.69	1.29
YaYi Llama 2 7B	58.42	49.97	1.29	79.09	1.05	68.70	1.19	50.63	1.29	54.27	1.28	61.42	1.25	47.45	1.29	64.79	1.23	50.03	1.29	53.94	1.28	62.34	1.25
TowerBase Llama 2 7B	57.78	49.17	1.29	77.23	1.08	69.82	1.18	50.76	1.29	52.88	1.28	58.31	1.27	48.38	1.29	67.04	1.21	50.36	1.29	53.14	1.28	58.50	1.27
TowerInstruct Llama 2 7B	59.24	49.31	1.29	80.87	1.01	71.61	1.16	50.69	1.29	52.95	1.28	59.56	1.26	48.71	1.29	69.36	1.19	51.49	1.29	54.14	1.28	63.00	1.24
EMMA-500 Llama 2 7B	66.38	66.25	1.22	76.44	1.09	70.02	1.18	64.73	1.23	64.92	1.23	68.63	1.19	57.91	1.27	68.50	1.20	64.73	1.23	64.66	1.23	63.40	1.24
Occiglot Mistral 7B v0.1	58.10	51.29	1.29	77.37	1.08	73.40	1.14	52.08	1.29	51.49	1.29	58.64	1.27	47.98	1.29	62.94	1.24	49.83	1.29	53.14	1.28	60.89	1.26
Occiglot Mistral 7B v0.1 Instruct	59.39	52.68	1.28	79.42	1.04	74.19	1.13	53.01	1.28	52.75	1.28	60.36	1.26	48.25	1.29	65.06	1.23	50.43	1.29	53.81	1.28	63.34	1.24
BLOOM 7B	59.30	58.57	1.27	70.55	1.17	66.18	1.22	57.25	1.27	60.42	1.26	64.53	1.23	48.91	1.29	52.75	1.28	53.94	1.28	57.31	1.27	61.88	1.25
BLOOMZ 7B	57.12	56.52	1.28	73.00	1.14	64.59	1.23	51.09	1.29	57.64	1.27	55.33	1.28	48.25	1.29	52.15	1.29	52.15	1.29	58.17	1.27	59.43	1.26
YaYi 7B	60.67	61.81	1.25	74.32	1.12	69.42	1.19	56.06	1.28	63.67	1.24	62.41	1.25	49.24	1.29	52.15	1.29	53.61	1.28	57.91	1.27	66.78	1.21
Aya 23 8B	60.93	62.61	1.25	74.59	1.12	67.31	1.21	50.96	1.29	64.26	1.23	66.64	1.21	47.72	1.29	68.70	1.19	50.36	1.29	54.00	1.28	63.14	1.24
Aya Expanse 8B	64.80	69.42	1.19	80.41	1.02	74.26	1.13	51.36	1.29	68.50	1.20	72.47	1.15	48.25	1.29	73.86	1.13	51.62	1.29	55.13	1.28	67.57	1.20
Gemma 7B	65.01	60.42	1.26	80.15	1.03	70.62	1.17	57.58	1.27	64.92	1.23	67.64	1.20	52.28	1.29	70.55	1.17	62.14	1.25	63.27	1.24	65.59	1.22
Gemma 2 9B	67.67	65.25	1.23	80.15	1.03	74.26	1.13	60.16	1.26	66.91	1.21	71.34	1.16	55.13	1.28	73.73	1.13	63.73	1.24	64.79	1.23	68.96	1.19
Qwen 1.5 7B	59.85	55.39	1.28	78.16	1.06	68.30	1.20	51.89	1.29	55.66	1.28	62.87	1.24	49.24	1.29	63.27	1.24	51.69	1.29	53.94	1.28	67.97	1.20
Qwen 2 7B	61.46	59.96	1.26	78.95	1.05	69.76	1.18	52.02	1.29	57.97	1.27	64.39	1.23	48.97	1.29	69.56	1.18	51.42	1.29	54.07	1.28	69.03	1.19
Qwen 2.5 7B	62.06	62.41	1.25	77.56	1.07	69.16	1.19	51.82	1.29	58.70	1.27	67.04	1.21	50.36	1.29	69.69	1.18	51.82	1.29	55.86	1.28	68.23	1.20
Marco-LLM GLO 7B	63.87	64.73	1.23	79.75	1.03	72.14	1.15	52.28	1.29	61.02	1.26	68.70	1.19	49.04	1.29	70.42	1.17	53.67	1.28	61.55	1.25	69.29	1.19
Llama 3 8B	63.41	58.64	1.27	78.69	1.05	70.62	1.17	55.79	1.28	62.81	1.24	65.92	1.22	51.09	1.29	68.76	1.19	56.39	1.28	63.00	1.24	65.78	1.22
Llama 3.1 8B	63.58	59.10	1.27	78.16	1.06	70.81	1.17	55.33	1.28	63.27	1.24	68.03	1.20	52.42	1.29	68.63	1.19	55.92	1.28	61.15	1.25	66.58	1.21
LLaMAX Llama 3 8B	64.31	59.96	1.26	76.97	1.08	69.09	1.19	59.23	1.26	62.87	1.24	66.78	1.21	55.33	1.28	68.30	1.20	60.69	1.26	63.60	1.24	64.59	1.23
LLaMAX Llama 3 8B Alpaca	68.27	64.73	1.23	83.19	0.96	73.00	1.14	59.96	1.26	67.50	1.21	72.40	1.15	57.31	1.27	72.80	1.15	64.00	1.24	65.19	1.23	70.88	1.17
EMMA-500 Llama 3 8B Mono	67.36	64.06	1.23	76.17	1.10	69.42	1.19	66.71	1.21	66.05	1.22	68.63	1.19	60.69	1.26	69.49	1.18	65.59	1.22	67.04	1.21	67.11	1.21
EMMA-500 Llama 3 8B Bi	68.35	66.38	1.22	75.78	1.10	71.28	1.16	66.64	1.21	66.91	1.21	69.56	1.18	63.00	1.24	70.95	1.17	67.37	1.21	67.31	1.21	66.71	1.21
EMMA-500 Llama 3.1 8B Mono	67.64	64.20	1.23	75.71	1.10	70.62	1.17	66.18	1.22	67.17	1.21	69.62	1.18	60.95	1.26	70.62	1.17	66.05	1.22	65.92	1.22	67.04	1.21
EMMA-500 Llama 3.1 8B Bi	68.47	66.18	1.22	75.12	1.11	71.34	1.16	66.45	1.22	67.77	1.20	70.68	1.17	62.54	1.25	71.41	1.16	67.50	1.21	66.58	1.21	67.64	1.20

Table 15: 0-shot results (Accuracy %) on XStoryCloze in all languages

long, MassiveSumm short, and XL-Sum. The prompt template for the three benchmarks is as follows:

```
Document: {text}
Based on the previous text, provide
    a brief single summary:
```

In order to try to get LLM to answer in the corresponding language of the document, we use Google Translate to translate the prompt into the corresponding language.

We evaluate the generated summaries using two widely recognized metrics: ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2019). ROUGE-L captures lexical overlap by measuring the longest common subsequence (LCS) between the reference and generated summaries. Precision and recall are calculated as the ratio of the LCS to the generated and reference summary lengths, respectively, with an F-score used for the final metric. BERTScore, on the other hand, measures semantic similarity by comparing contextual embeddings from pre-trained language models. Specifically, we

Model	Avg	High	Medium	Low
Llama 2 7B	40.19	45.26	37.72	34.97
Llama 2 7B Chat	38.58	42.77	36.75	33.87
CodeLlama 2 7B	40.19	46.27	37.29	33.86
LLaMAX Llama 2 7B	44.27	46.53	42.64	43.03
LLaMAX Llama 2 7B Alpaca	45.09	48.47	42.80	42.89
MaLA-500 Llama 2 10B v1	38.11	42.10	35.85	34.65
MaLA-500 Llama 2 10B v2	38.11	42.10	35.85	34.65
YaYi Llama 2 7B	41.28	47.32	38.41	34.94
TowerBase Llama 2 7B	39.84	46.08	36.33	34.39
TowerInstruct Llama 2 7B	40.36	47.07	36.92	33.79
EMMA-500 Llama 2 7B	45.14	46.09	44.40	44.71
Occiglot Mistral 7B v0.1	42.35	49.90	38.39	35.19
Occiglot Mistral 7B v0.1 Instruct	40.81	47.58	37.18	34.52
BLOOM 7B	41.60	45.13	39.69	38.38
BLOOMZ 7B	37.13	40.02	35.56	34.51
YaYi 7B	39.87	43.85	38.24	35.15
Aya 23 8B	43.12	48.51	41.95	34.67
Aya Expanse 8B	45.56	50.48	44.24	38.38
Gemma 7B	42.58	46.44	41.00	38.01
Gemma 2 9B	46.74	48.50	45.11	46.49
Qwen 1.5 7B	39.47	40.95	38.80	37.83
Qwen 2 7B	42.77	47.31	41.35	36.53
Qwen 2.5 7B	43.31	47.80	41.85	37.24
Marco-LLM GLO 7B	43.99	48.74	41.44	39.57
Llama 3 8B	44.97	48.82	43.84	39.56
Llama 3.1 8B	45.62	49.61	44.04	40.83
LLaMAX Llama 3 8B	44.13	46.30	42.83	42.41
LLaMAX Llama 3 8B Alpaca	45.08	48.18	42.82	43.40
EMMA-500 Llama 3 8B Mono	44.15	47.15	41.73	42.99
EMMA-500 Llama 3 8B Bi	45.15	47.20	43.64	44.07
EMMA-500 Llama 3.1 8B Mono	39.98	42.34	39.00	37.20
EMMA-500 Llama 3.1 8B Bi	44.67	46.79	42.88	44.00

Table 16: 0-shot results on XNLI (Accuracy %).

employ the `bert-base-multilingual-cased`[21] model, ensuring compatibility across multiple languages.

Our EMMA-500 models demonstrate consistent improvements over their Llama 3 and Llama 3.1 base models in zero-shot summarization across the MassiveSumm (long/short) and XL-Sum benchmarks (Tables 20, 21, 22). Notably, enhancements in semantic quality, measured by BERTScore, are frequently observed, especially on the MassiveSumm benchmarks. Nonetheless, Llama 3/3.1 base models do not obtain impressive summarization performance on these two subsets of MassiveSumm. On the XL-Sum benchmark, our EMMA-500 Llama 3/3.1 CPT models achieve better average performance than recent advances such as Aya 23 and Gemma 2.

D.7 Machine Reading Comprehension

We first evaluate on the BELEBELE dataset (Bandarkar et al., 2023), a comprehensive multilingual benchmark for machine reading comprehension, spanning 122 languages across both high- and low-resource categories. Notably, this benchmark demonstrates significant difficulty, with even the English subset posing substantial challenges to state-of-the-art language models like Llama 3. As shown in Table 23, which presents zero-shot performance across language resource groups, our continual pre-training approach on Llama 3 and 3.1, as well as LLaMAX CPT models, yields some degraded improvements over the base Llama model. However, when measured by the number of languages with improved performance, our EMMA-500 CPT is better than the base models. Contemporary models such as Aya Expanse, Llama 3.1, and Qwen 2 demonstrate more competent performance, suggesting advances in multilingual understanding and reasoning capacities. On the harder ARC-Multilingual set, the bilingual Llama-3 variant essentially ties the base model ($\approx 35\%$), while the others dip by up to one point. These results suggest that EMMA-style bilingual continual pre-training alone is insufficient for cross-language passage-question reasoning at the 8 B scale, the fundamental limitations of Llama models, and the average-based evaluation practice.

D.8 Math

We use the Multilingual Grade School Math Benchmark (MGSM) (Shi et al., 2022) to evaluate the mathematical reasoning in LLMs. As shown in Table 24, we evaluate model performance using 3-shot prompting with flexible answer matching. Our assessment employs two distinct prompting strategies: (1) Direct: direct question prompting, and (2) Chain-of-Thought (CoT): prompting accompanied by step-by-step reasoning examples (Wei et al., 2022). The results show that our CPT models, as well as LLaMAX CPT models, obtain decreased average performance on direct prompting. Our `EMMA-500 Llama 3.1 8B Mono` model slightly improves the base model. Notably, our EMMA-500 Llama 3/3.1 CPT models significantly improve the performance on low-resource languages for both direct and CoT prompting.

Tables 25 and 26 present the per-language performance across all the languages by direct and CoT prompting.

E Related Work

Multilingual Continual pre-training Multilingual language models have made significant progress in extending language understanding across diverse linguistic landscapes. Models such as mT5 (Xue et al., 2021) and BLOOM (Scao et al., 2022) have demonstrated strong multilingual capa-

[21] https://huggingface.co/google-bert/bert-base-multilingual-cased

Model	Avg	ar-acc	stderr	bg-acc	stderr	de-acc	stderr	el-acc	stderr	en-acc	stderr	es-acc	stderr	fr-acc	stderr	hi-acc	stderr	ru-acc	stderr	sw-acc	stderr	th-acc	stderr	tr-acc	stderr	ur-acc	stderr	vi-acc	stderr	zh-acc	stderr
Llama 2 7B	40.19	35.42	0.96	42.65	0.99	47.11	1.00	36.67	0.97	55.30	1.00	40.52	0.98	50.08	1.00	37.71	0.97	42.37	0.99	34.94	0.96	36.35	0.96	37.27	0.97	33.61	0.95	36.63	0.97	36.18	0.96
Llama 2 7B Chat	38.58	34.42	0.95	37.07	0.97	43.09	0.99	38.15	0.97	50.24	1.00	39.44	0.98	44.82	1.00	35.78	0.96	42.09	0.99	34.22	0.95	33.49	0.95	36.95	0.97	33.90	0.95	38.11	0.97	36.95	0.97
CodeLlama 2 7B	40.19	33.41	0.95	37.75	0.97	47.23	1.00	37.63	0.97	54.78	1.00	44.38	1.00	49.20	1.00	35.94	0.96	46.06	1.00	33.29	0.94	35.02	0.96	38.59	0.98	33.25	0.94	40.40	0.98	35.94	0.96
LLaMAX Llama 2 7B	44.27	33.78	0.95	46.83	1.00	48.96	1.00	42.57	0.99	54.90	1.00	47.59	1.00	47.79	1.00	45.50	1.00	45.54	1.00	43.05	0.99	41.85	0.99	43.29	0.99	44.18	1.00	43.86	0.99	34.38	0.95
LLaMAX Llama 2 7B Alpaca	45.09	34.42	0.95	46.39	1.00	49.76	1.00	43.41	0.99	58.11	0.99	48.96	1.00	51.97	1.00	45.62	1.00	46.27	1.00	43.57	0.99	40.80	0.99	43.86	0.99	44.30	1.00	43.09	0.99	35.78	0.96
MaLA-500 Llama 2 10B v1	38.11	35.94	0.96	41.20	0.99	47.51	1.00	34.46	0.95	56.18	0.99	34.10	0.95	47.59	1.00	33.65	0.95	33.94	0.95	35.22	0.96	33.69	0.95	33.82	0.95	35.02	0.96	36.02	0.96	33.25	0.94
MaLA-500 Llama 2 10B v2	38.11	35.94	0.96	41.20	0.99	47.51	1.00	34.46	0.95	56.18	0.99	34.10	0.95	47.59	1.00	33.65	0.95	33.94	0.95	35.22	0.96	33.69	0.95	33.82	0.95	35.02	0.96	36.02	0.96	33.25	0.94
YaYi Llama 2 7B	41.28	34.14	0.95	42.61	0.99	48.84	1.00	37.35	0.97	56.47	0.99	45.78	1.00	51.04	1.00	39.48	0.98	46.27	1.00	35.70	0.96	35.62	0.96	39.36	0.98	33.49	0.95	37.51	0.97	35.54	0.96
TowerBase Llama 2 7B	39.84	33.90	0.95	41.37	0.99	47.87	1.00	35.26	0.96	56.35	0.99	41.69	0.99	49.44	1.00	34.54	0.95	45.94	1.00	35.02	0.96	34.78	0.95	35.74	0.96	33.37	0.95	37.19	0.97	35.22	0.96
TowerInstruct Llama 2 7B	40.36	33.65	0.95	42.93	0.99	48.84	1.00	34.98	0.96	56.95	0.99	46.51	1.00	46.43	1.00	34.74	0.95	46.27	1.00	33.94	0.95	33.90	0.95	37.87	0.97	33.53	0.95	37.35	0.97	37.47	0.97
EMMA-500 Llama 2 7B	45.14	34.78	0.95	46.27	1.00	47.07	1.00	45.86	1.00	53.78	1.00	47.07	1.00	46.87	1.00	47.59	1.00	46.55	1.00	46.18	1.00	41.97	0.99	44.86	1.00	45.98	1.00	47.03	1.00	35.22	0.96
Occiglot Mistral 7B v0.1	42.35	33.86	0.95	41.37	0.99	51.77	1.00	37.71	0.97	55.86	1.00	51.65	1.00	51.93	1.00	35.74	0.96	47.63	1.00	34.70	0.95	37.39	0.97	43.01	0.99	33.49	0.95	38.63	0.98	40.56	0.98
Occiglot Mistral 7B v0.1 Instruct	40.81	34.38	0.95	38.84	0.98	50.84	1.00	40.00	0.98	55.66	1.00	48.63	1.00	51.69	1.00	34.30	0.95	40.04	0.98	33.17	0.94	36.35	0.96	37.99	0.97	34.06	0.95	37.59	0.97	38.63	0.98
BLOOM 7B	41.60	33.85	0.67	39.92	0.69	39.78	0.69	35.37	0.68	53.97	0.70	48.82	0.71	49.80	0.71	46.51	0.70	43.03	0.70	37.88	0.69	35.05	0.67	35.09	0.67	42.20	0.70	47.39	0.71	35.35	0.68
BLOOMZ 7B	37.13	32.69	0.94	34.02	0.95	41.69	0.99	35.82	0.96	46.87	1.00	36.02	0.96	43.05	0.99	40.40	0.98	37.47	0.97	33.61	0.95	33.09	0.94	33.73	0.95	36.83	0.97	36.67	0.97	35.02	0.96
YaYi 7B	39.87	39.80	0.98	35.78	0.96	42.61	0.99	36.47	0.96	50.52	1.00	47.71	1.00	48.19	1.00	40.04	0.98	39.12	0.98	34.06	0.95	34.34	0.95	33.17	0.94	37.07	0.97	44.18	1.00	34.94	0.96
Aya 23 8B	43.12	33.90	0.95	39.36	0.98	49.36	1.00	41.12	0.99	51.57	1.00	50.36	1.00	51.16	1.00	46.67	1.00	48.67	1.00	34.50	0.95	35.90	0.96	48.39	1.00	33.61	0.95	42.25	0.99	39.96	0.98
Aya Expanse 8B	45.56	34.10	0.95	41.89	0.99	51.04	1.00	42.17	0.99	53.86	1.00	47.35	1.00	53.65	1.00	48.31	1.00	51.00	1.00	37.03	0.97	41.04	0.99	48.84	1.00	37.07	0.97	50.12	1.00	45.98	1.00
Gemma 7B	42.58	33.49	0.95	43.49	0.99	48.63	1.00	38.11	0.97	52.05	1.00	44.14	1.00	49.76	1.00	44.34	1.00	47.39	1.00	40.64	0.98	37.95	0.97	43.25	0.99	35.42	0.96	43.29	0.99	36.67	0.97
Gemma 2 9B	46.74	34.18	0.95	49.52	1.00	51.37	1.00	43.25	0.99	53.45	1.00	51.41	1.00	52.29	1.00	47.31	1.00	49.56	1.00	45.58	1.00	49.88	1.00	50.72	1.00	44.02	1.00	45.70	1.00	32.93	0.94
Qwen 1.5 7B	39.47	34.34	0.95	40.92	0.99	42.77	0.99	36.18	0.96	49.08	1.00	37.87	0.97	43.13	0.99	38.15	0.97	38.88	0.98	35.42	0.96	44.22	1.00	38.84	0.98	33.86	0.95	44.38	1.00	33.98	0.95
Qwen 2 7B	42.77	33.69	0.95	45.46	1.00	48.19	1.00	36.83	0.97	54.26	1.00	47.23	1.00	51.41	1.00	44.98	1.00	47.39	1.00	37.31	0.97	38.27	0.97	43.53	0.99	34.02	0.95	43.57	0.99	35.38	0.96
Qwen 2.5 7B	43.31	34.38	0.95	42.85	0.99	47.31	1.00	41.45	0.99	53.65	1.00	48.92	1.00	51.57	1.00	42.49	0.99	47.71	1.00	34.22	0.95	43.33	0.99	42.57	0.99	34.18	0.95	47.39	1.00	37.63	0.97
Marco-LLM GLO 7B	43.99	33.45	0.95	45.50	1.00	50.40	1.00	37.35	0.97	53.82	1.00	50.52	1.00	51.57	1.00	42.13	0.99	46.63	1.00	34.70	0.95	44.62	1.00	44.82	1.00	39.40	0.98	45.38	1.00	39.52	0.98
Llama 3 8B	44.97	33.65	0.95	45.34	1.00	50.48	1.00	39.28	0.98	55.02	1.00	49.52	1.00	50.56	1.00	47.55	1.00	49.16	1.00	38.92	0.98	46.27	1.00	48.23	1.00	33.49	0.95	49.00	1.00	38.15	0.97
Llama 3.1 8B	45.62	33.86	0.95	45.58	1.00	51.45	1.00	38.96	0.98	55.22	1.00	50.28	1.00	51.77	1.00	49.40	1.00	49.16	1.00	39.36	0.98	48.07	1.00	49.32	1.00	35.06	0.96	47.11	1.00	39.76	0.98
LLaMAX Llama 3 8B	44.13	34.26	0.95	43.82	0.99	51.29	1.00	43.21	0.99	54.70	1.00	43.65	0.99	43.78	0.99	44.14	1.00	46.27	1.00	42.45	0.99	45.58	1.00	46.55	1.00	39.20	0.98	45.02	1.00	38.11	0.97
LLaMAX Llama 3 8B Alpaca	45.08	35.42	0.96	42.57	0.99	51.45	1.00	42.61	0.99	55.06	1.00	46.91	1.00	50.28	1.00	45.10	1.00	48.80	1.00	43.73	0.99	47.23	1.00	44.78	1.00	39.24	0.98	46.43	1.00	36.59	0.97
EMMA-500 Llama 3 8B Mono	44.15	33.01	0.94	46.06	1.00	48.59	1.00	37.91	0.97	56.39	0.99	47.43	1.00	47.95	1.00	45.78	1.00	47.79	1.00	46.14	1.00	41.00	0.99	44.50	1.00	41.81	0.99	43.09	0.99	34.74	0.95
EMMA-500 Llama 3 8B Bi	45.15	33.94	0.95	46.71	1.00	50.08	1.00	44.26	1.00	53.69	1.00	48.51	1.00	48.88	1.00	48.67	1.00	48.15	1.00	45.10	1.00	45.82	1.00	44.70	1.00	41.29	0.99	43.57	0.99	33.90	0.95
EMMA-500 Llama 3.1 8B Mono	39.98	34.66	0.95	41.33	0.99	42.61	0.99	35.62	0.96	47.99	1.00	42.33	0.99	42.41	0.99	41.16	0.99	44.58	1.00	38.47	0.98	37.19	0.97	40.56	0.98	35.94	0.96	40.68	0.98	34.10	0.95
EMMA-500 Llama 3.1 8B Bi	44.67	34.22	0.95	44.58	1.00	49.84	1.00	45.14	1.00	51.85	1.00	47.83	1.00	46.71	1.00	47.91	1.00	48.96	1.00	46.55	1.00	41.73	0.99	44.06	1.00	43.73	0.99	41.41	0.99	35.54	0.96

Table 17: 0-shot results (Accuracy %) on XNLI in all languages.

Model	Avg	High	Medium-high	Medium	Medium-low	Low
Llama 2 7B	12.93/30.32	19.98/38.72	12.90/30.87	13.13/31.00	15.02/33.00	13.69/31.26
Llama 2 7B Chat	12.28/31.72	19.12/39.31	11.96/31.89	12.24/32.08	14.32/34.15	13.05/32.61
CodeLlama 7B	10.82/28.57	17.77/37.45	10.72/29.09	10.93/29.19	12.77/31.13	11.52/29.47
LLaMAX Llama 2 7B	1.99/13.66	3.63/21.51	2.26/15.24	2.16/14.65	2.48/16.02	2.08/14.08
LLaMAX Llama 2 7B Alpaca	22.29/42.27	34.98/56.64	24.68/45.21	24.04/44.54	26.58/47.33	23.85/44.08
MaLA-500 10B V1	2.29/13.60	5.01/16.22	2.34/13.32	2.57/13.81	2.85/14.18	2.55/13.89
MaLA-500 10B V2	2.87/15.44	5.84/18.73	2.85/15.33	3.25/15.87	3.51/16.08	3.17/15.76
Yayi Llama 2 7B	12.98/31.38	19.77/39.36	12.93/31.97	13.13/32.03	15.02/33.96	13.76/32.30
Tower Base Llama 2 7B V0.1	13.74/31.47	21.68/40.54	13.81/32.18	13.97/32.28	16.05/34.31	14.50/32.44
Tower Instruct Llama 2 7B V0.2	4.81/25.43	9.75/33.70	5.29/25.73	5.18/25.97	5.85/27.70	5.24/26.32
EMMA-500 Llama 2 7B	25.37/45.78	34.84/56.32	26.72/47.21	26.93/47.47	28.89/49.76	27.34/48.00
Occiglot Mistral 7B	13.12/31.13	19.94/38.88	12.78/30.98	13.31/31.64	15.17/33.58	14.00/32.13
Occiglot Mistral 7B Instruct	11.61/31.65	16.92/39.16	11.27/31.61	11.80/32.21	13.20/33.98	12.31/32.61
BLOOM 7B1	9.57/27.84	15.83/36.50	12.72/32.18	11.77/31.02	10.88/29.80	9.96/28.51
BLOOMZ 7B1	20.22/34.74	31.74/46.29	26.57/41.03	24.76/39.25	22.39/37.03	21.06/35.63
Yayi 7B	4.82/21.36	5.04/23.09	6.25/24.58	5.69/23.82	5.20/22.08	4.72/21.25
Aya 23 8B	13.87/32.36	19.49/39.55	15.32/34.13	14.87/33.96	16.27/35.46	14.48/33.31
Aya Expanse 8B	13.12/36.86	17.35/44.58	14.77/39.30	14.54/39.15	14.81/39.83	13.61/37.85
Gemma 7B	23.79/43.68	36.37/57.74	26.72/47.06	26.34/46.77	27.78/48.45	25.47/45.65
Gemma 2 9B	23.15/38.87	34.42/51.93	23.37/38.63	24.85/40.38	27.22/43.58	25.19/41.36
Qwen 1.5 7B	15.58/35.87	24.44/46.60	16.96/38.19	16.97/38.06	18.17/39.11	16.42/37.01
Qwen 2 7B	17.39/37.61	28.37/50.71	19.49/40.81	19.53/40.66	20.66/41.61	18.40/38.90
Qwen 2.5 7B	18.95/38.89	29.32/51.08	21.04/41.75	21.21/41.80	22.11/42.65	19.97/40.11
Macro-LLM GLO 7B	25.17/44.57	35.96/56.86	28.34/48.18	28.15/47.90	28.54/48.47	26.38/45.89
Llama 3 8B	23.78/43.72	35.67/57.20	26.55/47.03	26.51/46.96	28.08/48.72	25.49/45.72
Llama 3.1 8B	24.19/44.10	36.13/57.57	26.76/47.27	26.73/47.25	28.40/49.05	25.91/46.13
LLaMAX Llama 3 8B	0.48/4.66	0.77/6.36	0.55/5.11	0.54/5.08	0.57/5.19	0.52/4.87
LLaMAX Llama 3 8B Alpaca	25.10/45.45	37.58/58.67	27.58/48.13	27.49/48.12	29.79/50.60	26.91/47.43
EMMA-500 Llama 3 8B Mono	30.38/50.86	39.52/60.25	31.90/52.47	32.44/53.10	34.13/54.81	32.55/53.16
EMMA-500 Llama 3 8B Bi	34.40/54.33	43.48/63.45	36.45/56.47	36.98/57.00	38.45/58.44	36.72/56.69
EMMA-500 Llama 3.1 8B Mono	27.57/48.12	37.34/58.82	29.18/49.85	29.50/50.33	31.20/52.20	29.51/50.34
EMMA-500 Llama 3.1 8B Bi	33.64/53.49	42.78/62.87	35.87/55.72	36.20/56.16	37.61/57.62	35.91/55.84

Table 18: 3-shot results (BLEU/chrF++) on FLORES-200, from all other languages to English (X-Eng)

bilities by pre-training on extensive multilingual corpora. However, its performance on low-resource languages remains limited. Recent developments such as Qwen 2 (Yang et al., 2024) and Llama 3 (Dubey et al., 2024) have further improved the efficiency and scalability of LLMs and shown stronger multilingual capabilities. Continual pre-training has emerged as an effective strategy for expanding language models' linguistic and domain coverage without requiring full retraining. Prior works such as MaLA-500 (Lin et al., 2024), LlaMAX (Lu et al., 2024), and EMMA-500 (Ji et al., 2024a) have demonstrated the feasibility of continual pre-training on diverse multilingual datasets. These approaches highlight the importance of incremental learning to incorporate new languages and improve cross-lingual transfer.

Despite these advancements, prior work has primarily focused on monolingual text pre-training, with limited emphasis on bilingual adaptation for improved cross-lingual transfer. Our work extends this line of research by introducing a bilingual translation corpus for continual pre-training and adapting Llama 3 and 3.1 (8B) to enhance bilingual and

Model	Avg	High	Medium-high	Medium	Medium-low	Low
Llama 2 7B	4.62/15.13	10.72/26.07	4.99/16.25	4.14/15.68	6.03/18.21	4.93/16.23
Llama 2 7B Chat	4.95/16.95	10.81/26.27	5.19/17.51	4.56/17.18	6.27/19.62	5.30/18.12
CodeLlama 7B	4.27/14.94	10.03/25.25	4.61/15.56	3.92/15.40	5.58/17.75	4.57/15.98
LLaMAX Llama 2 7B	0.80/7.42	2.08/13.88	1.14/8.98	0.92/8.14	1.04/8.88	0.86/7.92
LLaMAX Llama 2 7B Alpaca	12.51/28.35	26.39/45.94	16.08/32.23	14.07/30.87	16.00/34.14	13.36/30.45
MaLA-500 10B V1	0.60/6.08	1.62/10.25	0.69/6.15	0.57/6.26	0.80/7.13	0.65/6.64
MaLA-500 10B V2	0.54/6.38	1.37/10.37	0.59/6.45	0.46/6.42	0.67/7.31	0.55/6.93
Yayi Llama 2 7B	4.41/14.87	10.77/25.96	4.89/15.86	4.07/15.38	5.90/18.04	4.81/16.04
Tower Base Llama 2 7B V0.1	4.83/16.03	11.56/25.70	5.35/16.43	4.37/16.06	6.03/18.47	5.10/17.15
Tower Instruct Llama 2 7B V0.2	3.23/15.64	7.66/24.72	3.72/16.21	3.21/16.09	4.00/17.89	3.54/16.84
EMMA-500 Llama 2 7B	15.58/33.25	28.66/46.92	18.00/34.66	16.56/34.25	18.85/37.66	16.87/35.88
Occiglot Mistral 7B	4.32/16.10	10.97/25.72	5.13/16.65	4.36/16.57	5.51/18.54	4.69/17.23
Occiglot Mistral 7B Instruct	3.99/15.80	9.80/24.99	4.69/16.44	4.01/16.39	5.09/18.34	4.31/16.92
BLOOM 7B1	2.81/11.80	6.98/20.40	4.30/14.82	3.51/13.81	3.27/13.71	2.67/12.32
BLOOMZ 7B1	7.44/16.10	22.08/32.72	12.35/22.06	10.55/20.25	9.21/18.84	7.48/16.83
Yayi 7B	4.37/13.50	12.07/27.36	6.87/18.72	5.64/16.88	5.19/15.91	4.23/14.20
Aya 23 8B	6.46/16.15	10.76/23.27	8.16/18.42	6.53/17.12	7.75/18.86	6.41/16.89
Aya Expanse 8B	6.88/23.89	10.76/32.48	8.24/26.48	7.21/25.53	7.79/26.89	6.78/25.13
Gemma 7B	9.05/23.05	18.18/37.69	10.87/26.37	10.04/25.66	11.49/27.58	9.56/24.75
Gemma 2 9B	12.09/26.48	25.23/44.22	14.93/30.26	13.67/29.59	15.29/32.05	12.71/28.50
Qwen 1.5 7B	5.87/17.77	13.12/29.98	7.14/19.71	6.01/19.13	7.13/21.03	5.93/18.87
Qwen 2 7B	5.56/17.17	12.08/28.96	6.53/19.22	5.77/18.80	6.72/20.30	5.57/18.17
Qwen 2.5 7B	5.72/17.49	12.15/28.73	6.81/19.33	5.92/18.94	6.92/20.61	5.71/18.46
Macro-LLM GLO 7B	9.27/23.34	17.74/36.96	11.10/26.36	10.30/25.76	11.19/27.21	9.51/24.79
Llama 3 8B	9.93/24.08	20.79/40.03	11.76/26.96	10.97/26.59	12.58/28.89	10.51/25.84
Llama 3.1 8B	10.11/24.69	21.29/40.67	12.00/27.60	11.10/27.20	12.76/29.55	10.66/26.46
LLaMAX Llama 3 8B	0.45/4.65	1.11/7.77	0.64/5.45	0.56/5.19	0.57/5.28	0.49/4.93
LLaMAX Llama 3 8B Alpaca	11.64/26.86	25.20/45.23	14.71/30.78	13.25/29.63	14.99/32.76	12.44/28.94
EMMA-500 Llama 3 8B Mono	20.33/38.34	33.89/51.45	22.10/39.37	21.39/39.56	24.19/42.93	22.03/41.29
EMMA-500 Llama 3 8B Bi	24.02/42.15	38.25/55.19	25.73/43.49	25.32/43.74	28.32/47.00	25.99/45.22
EMMA-500 Llama 3.1 8B Mono	19.44/37.41	32.50/50.37	21.25/38.54	20.42/38.61	23.17/41.97	21.06/40.32
EMMA-500 Llama 3.1 8B Bi	23.86/42.07	37.81/54.81	25.76/43.65	25.26/43.74	28.15/46.84	25.80/45.07

Table 19: 3-shot results (BLEU/chrF++) on FLORES-200, from English to all other languages (Eng-X)

Model	Avg	High	Medium-high	Medium	Medium-low	Low
Llama 2 7B	4.74/63.89	3.88/61.92	3.88/61.92	3.40/62.75	4.78/63.95	4.78/63.95
Llama 2 7B Chat	4.73/63.52	3.88/61.57	3.88/61.57	3.42/62.34	4.76/63.62	4.76/63.62
CodeLlama 7B	5.63/64.51	4.45/60.73	4.45/60.73	4.11/63.72	5.58/64.54	5.58/64.54
LLaMAX Llama 2 7B	4.56/62.69	3.78/61.26	3.78/61.26	3.29/61.00	4.60/62.73	4.60/62.73
LLaMAX Llama 2 7B Alpaca	4.61/62.76	3.76/61.42	3.76/61.42	3.28/61.22	4.65/62.81	4.65/62.81
MaLA-500 10B V1	4.39/64.50	3.59/61.82	3.59/61.82	3.11/63.73	4.41/64.58	4.41/64.58
MaLA-500 10B V2	4.37/64.66	3.62/62.01	3.62/62.01	3.11/63.82	4.40/64.73	4.40/64.73
Yayi Llama 2 7B	4.98/64.17	4.05/62.17	4.05/62.17	3.69/63.15	5.02/64.21	5.02/64.21
Tower Base Llama 2 7B V0.1	4.81/64.51	3.88/62.39	3.88/62.39	3.43/63.49	4.85/64.59	4.85/64.59
Tower Instruct Llama 2 7B V0.2	4.82/64.61	3.84/62.52	3.84/62.52	3.40/63.49	4.85/64.67	4.85/64.67
EMMA-500 Llama 2 7B	4.79/63.80	3.81/61.99	3.81/61.99	3.43/62.66	4.82/63.88	4.82/63.88
Occiglot Mistral 7B	5.14/63.95	4.37/61.66	4.37/61.66	3.93/64.35	5.15/64.05	5.15/64.05
Occiglot Mistral 7B Instruct	5.16/63.50	4.57/61.79	4.57/61.79	4.19/63.86	5.19/63.53	5.19/63.53
BLOOM 7B1	4.88/64.36	4.00/62.34	4.00/62.34	3.76/64.65	4.92/64.45	4.92/64.45
BLOOMZ 7B1	2.91/57.20	2.12/58.71	2.12/58.71	1.89/57.54	2.94/57.17	2.94/57.17
Yayi 7B	4.95/64.24	4.26/61.87	4.26/61.87	3.81/64.25	4.99/64.32	4.99/64.32
Aya 23 8B	6.33/65.94	5.54/62.67	5.54/62.67	4.69/65.29	6.34/66.01	6.34/66.01
Aya Expanse 8B	7.44/67.66	6.29/64.39	6.29/64.39	5.65/66.65	7.48/67.76	7.48/67.76
Gemma 7B	6.18/62.14	5.25/60.19	5.25/60.19	4.80/62.14	6.23/62.19	6.23/62.19
Gemma 2 9B	5.86/59.70	4.96/57.51	4.96/57.51	4.81/59.83	5.89/59.74	5.89/59.74
Qwen 1.5 7B	6.09/59.19	5.42/58.28	5.42/58.28	5.07/60.73	6.09/59.14	6.09/59.14
Qwen 2 7B	6.65/56.31	6.47/54.91	6.47/54.91	5.43/54.74	6.64/56.16	6.64/56.16
Qwen 2.5 7B	7.44/61.62	6.88/58.85	6.88/58.85	6.19/59.98	7.50/61.59	7.50/61.59
Macro-LLM GLO 7B	6.57/57.15	5.84/53.04	5.84/53.04	5.26/55.92	6.60/57.11	6.60/57.11
Llama 3 8B	5.10/55.58	4.15/54.48	4.15/54.48	4.04/57.63	5.13/55.65	5.13/55.65
Llama 3.1 8B	5.41/56.09	4.36/55.24	4.36/55.24	4.34/57.85	5.45/56.19	5.45/56.19
LLaMAX Llama 3 8B	6.00/65.90	4.74/62.84	4.74/62.84	4.57/65.35	6.03/66.00	6.03/66.00
LLaMAX Llama 3 8B Alpaca	7.62/67.64	6.13/64.77	6.13/64.77	5.92/66.58	7.69/67.77	7.69/67.77
EMMA-500 Llama 3 8B Mono	5.38/61.06	4.61/57.84	4.61/57.84	4.42/60.75	5.41/61.08	5.41/61.08
EMMA-500 Llama 3 8B Bi	5.57/59.23	4.85/57.46	4.85/57.46	4.67/60.13	5.62/59.30	5.62/59.30
EMMA-500 Llama 3.1 8B Mono	5.44/61.89	4.70/59.04	4.70/59.04	4.49/61.99	5.48/61.98	5.48/61.98
EMMA-500 Llama 3.1 8B Bi	4.78/58.47	3.70/53.79	3.70/53.79	3.90/58.36	4.81/58.48	4.81/58.48

Table 20: Zero-short performance of MassiveSumm long set (ROUGE-L/BERTScore). Our EMMA-500 models demonstrate consistent improvements over their Llama 3 and Llama 3.1 base models.

multilingual generalization, particularly for under-represented languages.

Training on Bilingual Translation Data Parallel texts have been used for pre-training LLMs. PolyLM (Wei et al., 2023) trained on 1 billion parallel multilingual data (0.16% of pre-training corpora). Poro (Luukkonen et al., 2024) trained on 8 billion English-Finnish cross-lingual texts (slightly

Model	Avg	High	Medium-high	Medium	Medium-low	Low
Llama 2 7B	7.85/65.35	6.64/62.55	6.23/62.78	5.92/64.92	7.79/65.43	7.88/65.33
Llama 2 7B Chat	9.76/67.01	8.88/64.52	8.29/64.64	7.59/66.03	9.83/67.08	9.87/66.98
CodeLlama 7B	7.59/64.83	6.19/61.81	5.88/62.10	5.54/64.49	7.52/65.03	7.60/64.93
LLaMAX Llama 2 7B	5.22/63.06	4.47/59.76	4.10/59.92	3.46/62.60	5.16/63.15	5.26/63.10
LLaMAX Llama 2 7B Alpaca	10.71/67.92	9.52/64.83	8.93/65.09	8.12/66.93	10.73/67.99	10.81/67.90
MaLA-500 10B V1	4.97/63.51	4.34/61.56	4.00/61.65	3.53/63.32	5.04/63.76	5.13/63.74
MaLA-500 10B V2	5.02/63.75	4.34/61.61	4.00/61.69	3.55/63.46	5.09/63.97	5.18/63.95
Yayi Llama 2 7B	7.80/65.24	6.91/63.23	6.46/63.33	5.96/64.80	7.84/65.34	7.87/65.21
Tower Base Llama 2 7B V0.1	8.11/65.53	6.70/62.59	6.26/62.78	5.94/64.96	8.04/65.58	8.12/65.48
Tower Instruct Llama 2 7B V0.2	10.14/67.76	8.88/65.06	8.24/65.12	7.44/66.81	10.15/67.86	10.18/67.71
EMMA-500 Llama 2 7B	8.32/65.14	6.85/62.52	6.41/62.56	6.13/64.26	8.30/65.25	8.43/65.16
Occiglot Mistral 7B	8.16/63.65	6.39/61.67	6.05/61.74	5.79/64.17	8.15/64.00	8.26/63.96
Occiglot Mistral 7B Instruct	7.82/63.79	6.72/62.31	6.37/62.13	6.05/64.25	7.72/63.74	7.84/63.71
BLOOM 7B1	6.79/62.30	5.71/58.74	5.34/58.80	4.97/61.23	6.73/62.31	6.82/62.26
BLOOMZ 7B1	3.28/29.75	3.72/39.04	3.41/36.51	2.78/32.31	3.16/29.30	3.20/29.11
Yayi 7B	8.28/65.44	7.73/65.10	7.25/64.79	6.62/65.09	8.30/65.66	8.35/65.42
Aya 23 8B	8.43/65.85	7.49/63.54	6.99/63.73	6.14/65.59	8.47/66.04	8.47/65.85
Aya Expanse 8B	9.24/67.68	8.49/65.77	7.96/65.95	7.38/67.20	9.32/67.88	9.39/67.72
Gemma 7B	8.35/62.25	7.56/60.19	7.09/60.45	6.39/62.35	8.35/62.28	8.47/62.29
Gemma 2 9B	7.86/58.11	6.36/55.20	5.86/53.92	5.95/57.87	7.85/57.99	7.97/58.11
Qwen 1.5 7B	8.49/62.70	7.88/61.33	7.40/61.00	6.87/63.25	8.44/62.60	8.50/62.57
Qwen 2 7B	8.63/56.14	7.65/53.08	7.17/53.36	6.33/55.72	8.49/55.57	8.54/55.69
Qwen 2.5 7B	9.04/58.91	9.49/60.70	8.85/60.34	7.64/60.70	9.03/58.90	9.04/58.56
Macro-LLM GLO 7B	8.10/57.45	8.09/56.96	7.51/57.15	6.87/60.07	8.06/57.25	8.10/57.29
Llama 3 8B	6.44/50.98	4.78/48.31	4.53/48.77	4.68/52.54	6.29/50.70	6.45/50.91
Llama 3.1 8B	6.77/54.96	5.31/52.26	5.02/53.04	4.91/55.77	6.67/54.79	6.83/54.97
LLaMAX Llama 3 8B	8.77/66.46	7.05/63.80	6.60/63.77	6.27/65.97	8.65/66.58	8.76/66.47
LLaMAX Llama 3 8B Alpaca	12.44/68.95	10.27/67.22	9.54/67.08	9.42/68.35	12.44/69.09	12.48/68.94
EMMA-500 Llama 3 8B Mono	7.18/63.46	6.39/61.99	5.97/62.06	5.68/64.19	7.19/63.69	7.24/63.50
EMMA-500 Llama 3 8B Bi	7.74/63.18	6.49/60.71	6.01/60.71	5.72/63.07	7.76/63.28	7.82/63.26
EMMA-500 Llama 3.1 8B Mono	7.21/63.36	6.68/63.09	6.15/61.81	5.67/62.64	7.25/63.45	7.34/63.46
EMMA-500 Llama 3.1 8B Bi	6.67/61.64	5.61/58.37	5.21/58.16	5.01/61.53	6.67/61.59	6.75/61.64

Table 21: Zero-short performance of MassiveSumm short set (ROUGE-L/BERTScore). Our EMMA-500 models demonstrate consistent improvements over their Llama 3 and Llama 3.1 base models.

Model	Avg	High	Medium-high	Medium	Medium-low	Low
Llama 2 7B	7.11/66.52	9.75/64.83	9.52/65.35	8.34/65.08	7.62/65.58	7.73/65.39
Llama 2 7B Chat	8.84/68.44	12.91/67.19	12.32/67.84	10.47/67.45	9.54/67.69	9.67/67.51
CodeLlama 7B	7.15/65.74	9.89/62.99	9.66/64.20	8.38/63.72	7.71/64.74	7.83/64.68
LLaMAX Llama 2 7B	5.29/64.59	6.86/61.74	6.66/62.29	5.99/61.96	5.71/63.24	5.79/63.17
LLaMAX Llama 2 7B Alpaca	10.11/69.24	14.44/67.88	14.31/68.35	12.42/68.26	11.12/68.38	11.19/68.19
MaLA-500 10B V1	5.45/63.96	7.98/60.89	7.40/61.16	6.38/61.14	5.88/62.48	5.94/62.55
MaLA-500 10B V2	5.44/64.28	7.80/61.21	7.28/61.62	6.34/61.51	5.86/62.71	5.91/62.79
Yayi Llama 2 7B	7.98/67.21	11.64/66.06	10.91/66.54	9.41/66.17	8.59/66.63	8.71/66.32
Tower Base Llama 2 7B V0.1	7.65/67.09	10.64/65.26	10.45/65.62	8.97/65.39	8.21/66.12	8.30/65.97
Tower Instruct Llama 2 7B V0.2	8.89/68.46	12.81/67.15	12.00/67.56	10.30/67.06	9.53/67.72	9.65/67.52
EMMA-500 Llama 2 7B	8.58/67.20	11.61/66.17	11.75/66.93	10.29/67.92	9.41/67.80	9.52/67.68
Occiglot Mistral 7B	7.33/66.20	10.38/63.81	10.17/64.44	8.86/64.39	7.93/65.00	8.02/64.88
Occiglot Mistral 7B Instruct	8.31/66.96	12.53/65.19	11.44/65.98	10.08/65.16	9.04/66.17	9.11/66.04
BLOOM 7B1	6.99/64.78	8.83/61.91	8.62/63.21	7.64/62.69	7.43/63.72	7.55/63.43
BLOOMZ 7B1	11.15/69.82	20.33/67.38	19.39/68.58	15.83/68.77	12.52/69.37	12.50/69.21
Yayi 7B	12.06/69.74	19.99/67.84	19.25/68.82	16.07/68.85	13.42/69.14	13.37/68.97
Aya 23 8B	8.68/66.79	12.91/66.50	11.90/66.45	9.92/67.34	9.25/67.11	9.43/66.91
Aya Expanse 8B	10.51/68.73	14.88/67.38	14.09/67.96	12.58/69.15	11.24/69.07	11.32/68.88
Gemma 7B	6.70/64.52	9.16/62.37	9.16/62.76	8.17/64.49	7.33/64.93	7.30/64.67
Gemma 2 9B	7.38/65.45	9.82/61.04	9.83/62.66	8.91/62.78	8.01/64.14	8.07/64.11
Qwen 1.5 7B	9.58/69.13	14.33/67.56	13.58/68.14	12.04/68.02	10.42/68.34	10.44/68.16
Qwen 2 7B	10.18/69.35	15.67/68.37	14.66/68.66	12.70/68.42	11.12/68.58	11.10/68.39
Qwen 2.5 7B	10.41/69.69	15.68/69.57	14.96/69.69	12.95/70.62	11.38/70.19	11.31/69.93
Macro-LLM GLO 7B	11.46/70.41	17.22/70.74	16.65/70.83	14.31/71.32	12.65/71.03	12.62/70.76
Llama 3 8B	8.47/67.08	10.96/64.16	11.24/65.16	10.12/65.04	9.29/65.82	9.31/65.79
Llama 3.1 8B	8.57/66.97	10.97/63.60	11.19/64.55	10.14/64.81	9.45/65.66	9.43/65.66
LLaMAX Llama 3 8B	8.28/66.72	10.85/64.80	10.97/65.33	9.71/66.92	9.04/67.25	9.07/66.99
LLaMAX Llama 3 8B Alpaca	11.39/69.99	15.45/69.66	15.82/70.20	13.36/70.75	12.44/70.53	12.42/70.34
EMMA-500 Llama 3 8B Mono	9.11/66.12	12.96/64.51	13.25/65.23	11.30/66.81	10.09/66.83	9.96/66.50
EMMA-500 Llama 3 8B Bi	9.10/66.72	13.30/65.28	13.11/65.91	11.17/67.34	10.10/67.46	9.94/67.12
EMMA-500 Llama 3.1 8B Mono	9.66/67.21	12.88/65.35	13.18/66.30	11.53/67.78	10.63/67.93	10.58/67.66
EMMA-500 Llama 3.1 8B Bi	8.56/65.90	12.19/64.10	11.75/64.40	10.51/66.43	9.51/66.53	9.47/66.31

Table 22: Zero-short performance of XL-Sum (ROUGE-L/BERTScore). Our EMMA-500 models demonstrate consistent improvements over their Llama 3 and Llama 3.1 base models.

under 1% of pre-training corpora). Li et al. (2024b) trained small language models, i.e., BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), and BART (Lewis, 2019), in a controlled setting with multilingual and bilingual texts using different learning objectives. In continual pre-training

Model	BELEBELE						ARC Multilingual			
	Avg	High	Med-high	Medium	Med-low	Low	Avg	High	Medium	Low
Llama 2 7B	26.27	26.76	26.35	26.07	26.41	26.27	27.56	33.12	27.31	21.02
Llama 2 7B Chat	29.05	31.84	29.97	28.95	29.47	29.09	28.02	33.69	27.79	21.29
CodeLlama 2 7B	27.38	27.37	27.33	27.30	27.30	27.38	25.23	28.86	24.64	21.65
LLaMAX Llama 2 7B	23.09	23.23	23.15	23.07	23.10	23.08	26.09	30.00	25.92	21.48
LLaMAX Llama 2 7B Alpaca	24.48	25.46	24.82	24.41	24.60	24.49	31.06	36.89	31.85	22.49
MaLA-500 Llama 2 10B v1	22.96	23.02	22.98	22.97	22.98	22.97	21.16	21.92	20.48	21.32
MaLA-500 Llama 2 10B v2	22.96	23.02	22.98	22.97	22.98	22.97	21.16	21.92	20.48	21.32
YaYi Llama 2 7B	28.32	29.64	28.67	28.11	28.37	28.26	28.40	34.30	28.35	21.11
TowerBase Llama 2 7B	26.36	27.43	26.85	26.29	26.48	26.34	27.94	35.32	26.82	20.51
TowerInstruct Llama 2 7B	27.93	29.88	28.51	27.57	28.19	27.92	30.10	38.88	28.85	21.16
EMMA-500 Llama 2 7B	26.75	28.32	28.18	27.58	27.14	26.94	29.53	34.10	29.82	23.34
Occiglot Mistral 7B v0.1	30.16	32.25	30.94	30.02	30.40	30.15	29.77	38.39	28.51	21.03
Occiglot Mistral 7B v0.1 Instruct	32.05	34.14	32.62	31.74	32.40	32.08	30.88	40.29	29.65	21.13
BLOOM 7B	24.11	24.25	24.52	24.12	24.11	24.08	23.65	26.27	22.72	21.89
BLOOMZ 7B	39.32	45.43	43.67	41.51	40.08	39.51	23.95	26.94	22.74	22.18
YaYi 7B	37.97	44.37	42.71	40.49	38.72	38.09	24.44	27.96	23.29	21.91
Aya 23 8B	40.08	43.85	41.71	39.22	40.93	39.81	31.08	40.05	30.02	21.61
Aya Expanse 8B	46.98	52.22	48.99	46.57	48.36	46.93	36.56	47.87	36.15	23.09
Gemma 7B	43.37	52.63	47.83	44.82	45.43	43.94	38.68	46.46	40.47	26.06
Gemma 2 9B	54.49	64.10	58.90	55.83	56.85	55.05	44.15	54.59	46.18	27.82
Qwen 1.5 7B	41.83	48.86	44.79	42.18	43.00	41.78	28.93	35.55	28.14	21.92
Qwen 2 7B	49.31	57.62	52.20	50.04	51.16	49.48	33.82	43.88	32.64	23.17
Qwen 2.5 7B	54.11	63.47	57.91	55.47	56.24	54.30	35.30	46.43	34.31	23.00
Marco-LLM GLO 7B	53.95	63.54	58.67	55.50	56.20	54.31	36.34	46.72	35.88	24.11
Llama 3 8B	40.73	46.07	42.92	41.03	41.87	40.88	34.80	42.43	35.53	24.06
Llama 3.1 8B	45.19	52.50	48.01	45.65	46.74	45.34	34.93	42.43	35.89	24.00
LLaMAX Llama 3 8B	36.96	40.06	37.92	36.78	37.68	37.09	33.54	39.81	34.84	23.59
LLaMAX Llama 3 8B Alpaca	39.41	44.06	41.10	39.52	40.53	39.61	34.53	41.34	35.56	24.34
EMMA-500 Llama 3 8B Mono	39.73	43.65	41.40	39.96	40.76	40.01	33.22	38.32	34.56	24.67
EMMA-500 Llama 3 8B Bi	39.84	43.51	41.09	40.05	40.74	39.99	34.84	40.37	36.45	25.31
EMMA-500 Llama 3.1 8B Mono	38.86	41.68	39.34	38.49	39.62	38.94	34.00	39.38	35.39	25.01
EMMA-500 Llama 3.1 8B Bi	37.00	40.59	38.36	37.21	37.94	37.25	34.59	39.85	36.21	25.38

Table 23: 0-shot results (Accuracy %) on BELEBELE, and 5-shot results (Accuracy %) on the multilingual ARC.

Model	Direct				CoT			
	Avg	High	Medium	Low	Avg	High	Medium	Low
Llama 2 7B	6.69	8.07	2.13	1.20	6.36	7.60	2.13	0.80
Llama 2 7B Chat	10.22	13.73	2.13	0.80	10.91	13.53	2.80	1.60
CodeLlama 2 7B	5.93	7.07	2.93	1.20	6.64	8.73	2.67	2.00
LLaMAX Llama 2 7B	3.35	4.00	2.00	0.80	3.62	4.33	2.27	2.40
LLaMAX Llama 2 7B Alpaca	5.05	5.20	4.00	1.60	6.35	8.07	4.13	0.80
MaLA-500 Llama 2 10B v1	0.91	1.33	0.27	0.00	0.73	1.27	0.00	0.00
MaLA-500 Llama 2 10B v2	0.91	1.33	0.27	0.00	0.73	1.27	0.00	0.00
YaYi Llama 2 7B	7.09	9.47	1.47	1.20	7.22	8.73	2.27	1.20
TowerBase Llama 2 7B	6.15	8.33	1.73	0.80	6.16	8.60	2.40	0.80
TowerInstruct Llama 2 7B	7.24	9.53	1.73	2.00	8.24	10.47	1.87	1.20
EMMA-500 Llama 2 7B	17.02	19.20	11.87	2.40	18.09	20.00	13.20	2.80
Occiglot Mistral 7B v0.1	13.31	16.87	4.53	1.60	14.07	18.80	3.60	1.20
Occiglot Mistral 7B v0.1 Instruct	22.76	29.80	7.47	2.80	22.16	30.40	7.87	2.80
BLOOM 7B	2.87	2.60	2.80	3.60	2.29	2.20	1.47	2.00
BLOOMZ 7B	2.55	2.67	2.40	1.20	2.15	1.67	3.07	2.00
YaYi 7B	2.76	2.93	1.47	2.40	3.02	2.93	2.40	2.00
Aya 23 8B	22.29	30.67	3.47	0.80	24.71	35.07	5.47	2.40
Aya Expanse 8B	43.02	55.47	18.93	5.20	41.45	54.67	18.53	5.20
Gemma 7B	38.22	36.60	38.27	27.20	35.78	34.67	37.07	26.80
Gemma 2 9B	32.95	28.00	35.73	30.80	44.69	36.07	52.00	47.20
Qwen 1.5 7B	31.56	40.00	16.00	4.00	30.36	40.60	14.80	2.40
Qwen 2 7B	48.95	54.40	38.80	14.80	51.47	58.93	39.07	14.40
Qwen 2.5 7B	53.78	65.33	36.13	8.00	55.60	68.87	36.40	8.80
Marco-LLM GLO 7B	51.85	63.80	32.80	11.60	52.02	62.93	34.93	11.60
Llama 3 8B	27.45	27.87	26.13	5.60	28.13	28.53	26.67	5.20
Llama 3.1 8B	28.36	29.00	26.13	4.40	27.31	27.27	25.47	8.40
LLaMAX Llama 3 8B	20.80	22.73	20.40	6.80	19.96	20.93	22.40	3.60
LLaMAX Llama 3 8B Alpaca	14.18	14.87	11.87	6.00	17.16	20.87	12.27	4.00
EMMA-500 Llama 3 8B Mono	23.53	22.33	25.60	12.00	25.33	24.67	27.73	13.60
EMMA-500 Llama 3 8B Bi	23.49	23.00	24.67	10.80	26.29	24.87	30.67	11.60
EMMA-500 Llama 3.1 8B Mono	24.95	23.67	27.20	11.20	27.35	26.60	30.27	10.80
EMMA-500 Llama 3.1 8B Bi	23.85	23.67	25.33	11.20	25.76	26.27	27.87	12.80

Table 24: 3-shot results (Accuracy %) on MGSM obtained by direct and CoT prompting. Our EMMA-500 CPT models significantly improve the performance on low-resource languages.

or fine-tuning, bilingual texts are also widely used. Ji et al. (2024b) showed continual pre-training of multilingual BART (Tang et al., 2020) with machine translation failed enhance cross-lingual representation learning. Xu et al. (2024) fine-tuned LLMs with monolingual data and parallel data for

Model	Avg	bn	bn-stderr	de	de-stderr	en	en-stderr	es	es-stderr	fr	fr-stderr	ja	ja-stderr	ru	ru-stderr	sw	sw-stderr	te	te-stderr	th	th-stderr	zh	zh-stderr
Llama 2 7B	6.69	2.80	1.05	8.00	1.72	17.60	2.41	11.20	2.00	12.00	2.06	2.40	0.97	8.00	1.72	2.80	1.05	1.20	0.69	0.80	0.56	6.80	1.60
Llama 2 7B Chat	10.22	2.80	1.05	16.80	2.37	22.80	2.66	19.60	2.52	19.20	2.50	2.40	0.97	14.40	2.22	0.80	0.56	0.80	0.56	2.80	1.05	10.00	1.90
CodeLlama 2 7B	5.93	1.60	0.80	8.80	1.80	12.80	2.12	8.80	1.80	10.40	1.93	5.60	1.46	6.00	1.51	2.00	0.89	1.20	0.69	5.20	1.41	2.80	1.05
LLaMAX Llama 2 7B	3.35	2.80	1.05	3.60	1.18	6.00	1.51	2.00	0.89	7.20	1.64	3.20	1.12	2.40	0.97	1.60	0.80	0.80	0.56	1.60	0.80	5.60	1.46
LLaMAX Llama 2 7B Alpaca	5.05	4.00	1.24	3.60	1.18	10.80	1.97	6.00	1.51	6.40	1.55	3.20	1.12	4.40	1.30	4.80	1.35	1.60	0.80	3.20	1.12	7.60	1.68
MaLA-500 Llama 2 10B v1	0.91	0.00	0.00	0.00	0.00	1.20	0.69	2.00	0.89	2.40	0.97	1.20	0.69	2.00	0.89	0.40	0.40	0.00	0.00	0.40	0.40	0.40	0.40
MaLA-500 Llama 2 10B v2	0.91	0.00	0.00	0.00	0.00	1.20	0.69	2.00	0.89	2.40	0.97	1.20	0.69	2.00	0.89	0.40	0.40	0.00	0.00	0.40	0.40	0.40	0.40
YaYi Llama 2 7B	7.09	3.20	1.12	8.40	1.76	15.60	2.30	16.00	2.32	10.40	1.93	5.20	1.41	5.60	1.46	0.80	0.56	1.20	0.69	0.40	0.40	11.20	2.00
TowerBase Llama 2 7B	6.15	2.40	0.97	8.40	1.76	11.60	2.03	9.20	1.83	8.80	1.80	4.80	1.35	10.00	1.90	1.20	0.69	0.80	0.56	1.60	0.80	8.80	1.80
TowerInstruct Llama 2 7B	7.24	1.60	0.80	10.00	1.90	15.20	2.28	15.60	2.30	12.80	2.12	1.60	0.80	10.00	1.90	1.60	0.80	2.00	0.89	2.00	0.89	7.20	1.64
EMMA-500 Llama 2 7B	17.02	8.80	1.80	23.20	2.68	34.00	3.00	28.00	2.85	25.60	2.77	9.20	1.83	22.80	2.66	16.80	2.37	2.40	0.97	10.00	1.90	6.40	1.55
Occiglot Mistral 7B v0.1	13.31	3.20	1.12	21.20	2.59	30.00	2.90	27.20	2.82	21.60	2.61	6.40	1.55	15.20	2.28	2.40	0.97	1.60	0.80	8.00	1.72	9.60	1.87
Occiglot Mistral 7B v0.1 Instruct	22.76	4.80	1.35	34.00	3.00	46.40	3.16	40.00	3.10	31.60	2.95	18.40	2.46	23.60	2.69	6.40	1.55	2.80	1.05	11.20	2.00	31.20	2.94
BLOOM 7B	2.87	2.40	0.97	1.60	0.80	4.00	1.24	3.60	1.18	1.20	0.69	2.00	0.89	3.60	1.18	4.00	1.24	3.60	1.18	2.00	0.89	3.60	1.18
BLOOMZ 7B	2.55	3.20	1.12	1.60	0.80	3.60	1.18	2.40	0.97	3.20	1.12	2.80	1.05	3.60	1.18	2.80	1.05	1.20	0.69	1.20	0.69	2.40	0.97
YaYi 7B	2.76	2.40	0.97	2.00	0.89	6.00	1.51	2.40	0.97	5.60	1.46	1.60	0.80	1.20	0.69	1.20	0.69	2.40	0.97	0.80	0.56	4.80	1.35
Aya 23 8B	22.29	3.20	1.12	37.20	3.06	50.00	3.17	41.20	3.12	39.20	3.09	27.60	2.83	34.00	3.00	2.80	1.05	0.80	0.56	4.40	1.30	4.80	1.35
Aya Expanse 8B	43.02	22.40	2.64	69.20	2.93	78.40	2.61	72.00	2.85	63.60	3.05	50.80	3.17	69.60	2.92	12.80	2.12	5.20	1.41	21.60	2.61	7.60	1.68
Gemma 7B	38.22	34.40	3.01	44.80	3.15	58.80	3.12	48.00	3.17	39.60	3.10	16.80	2.37	41.20	3.12	37.60	3.07	27.20	2.82	42.80	3.14	29.20	2.88
Gemma 2 9B	32.95	29.60	2.89	40.40	3.11	56.40	3.14	50.80	3.17	36.00	3.04	1.20	0.69	35.20	3.03	37.60	3.07	30.80	2.93	40.00	3.10	4.40	1.30
Qwen 1.5 7B	31.56	12.40	2.09	44.00	3.15	55.20	3.15	48.40	3.17	45.20	3.15	16.80	2.37	43.20	3.14	7.20	1.64	4.00	1.24	28.40	2.86	42.40	3.13
Qwen 2 7B	48.95	44.40	3.15	65.60	3.01	80.80	2.50	76.00	2.71	69.60	2.92	1.60	0.80	67.20	2.98	16.00	2.32	14.80	2.25	56.00	3.15	46.40	3.16
Qwen 2.5 7B	53.78	26.80	2.81	63.20	3.06	83.20	2.37	74.00	2.78	66.40	2.99	52.40	3.16	72.80	2.82	20.00	2.53	8.00	1.72	61.60	3.08	63.20	3.06
Marco-LLM GLO 7B	51.85	31.60	2.95	64.40	3.03	77.60	2.64	71.20	2.87	63.20	3.06	50.40	3.17	63.60	3.05	18.00	2.43	11.60	2.03	48.80	3.17	70.00	2.90
Llama 3 8B	27.45	17.60	2.41	39.60	3.10	50.80	3.17	47.20	3.16	36.40	3.05	3.60	1.18	37.60	3.07	24.00	2.71	5.60	1.46	36.80	3.06	2.80	1.05
Llama 3.1 8B	28.36	20.00	2.53	41.20	3.12	55.20	3.15	48.40	3.17	38.40	3.08	3.60	1.18	40.00	3.10	20.80	2.57	4.40	1.30	37.60	3.07	2.40	0.97
LLaMAX Llama 3 8B	20.80	13.20	2.15	20.40	2.55	24.40	2.72	26.00	2.78	22.00	2.63	20.80	2.57	22.80	2.66	21.20	2.59	6.80	1.60	26.80	2.81	24.40	2.72
LLaMAX Llama 3 8B Alpaca	14.18	12.00	2.06	17.60	2.41	25.20	2.75	17.20	2.39	17.20	2.39	11.20	2.00	15.60	2.30	10.80	1.97	6.00	1.51	12.80	2.12	10.40	1.93
EMMA-500 Llama 3 8B Mono	23.53	20.00	2.53	30.00	2.90	36.00	3.04	34.80	3.02	30.00	2.90	3.20	1.12	33.20	2.98	33.20	2.98	12.00	2.06	23.60	2.69	2.80	1.05
EMMA-500 Llama 3 8B Bi	23.49	21.20	2.59	30.80	2.93	35.60	3.03	36.80	3.06	34.00	3.00	2.80	1.05	30.40	2.92	30.00	2.90	10.80	1.97	22.80	2.66	3.20	1.12
EMMA-500 Llama 3.1 8B Mono	24.95	21.20	2.59	34.80	3.02	39.60	3.10	35.60	3.03	33.20	2.98	1.20	0.69	34.80	3.02	36.80	3.06	11.20	2.00	23.60	2.69	2.40	0.97
EMMA-500 Llama 3.1 8B Bi	23.85	20.80	2.57	38.00	3.08	33.20	2.98	39.60	3.10	32.40	2.97	2.00	0.89	28.40	2.86	27.20	2.82	11.20	2.00	28.00	2.85	1.60	0.80

Table 25: 3-shot results (Accuracy %) on MGSM in all languages by direct prompting and flexible matching.

Model	Avg	bn	bn-stderr	de	de-stderr	en	en-stderr	es	es-stderr	fr	fr-stderr	ja	ja-stderr	ru	ru-stderr	sw	sw-stderr	te	te-stderr	th	th-stderr	zh	zh-stderr
Llama 2 7B	6.36	2.00	0.89	7.60	1.68	16.00	2.32	12.40	2.09	9.20	1.83	3.60	1.18	8.00	1.72	2.00	0.89	0.80	0.56	1.60	0.80	8.00	1.72
Llama 2 7B Chat	10.91	2.40	0.97	17.60	2.41	27.20	2.82	19.20	2.50	18.80	2.48	4.00	1.24	15.20	2.28	2.00	0.89	0.80	0.56	2.80	1.05	11.60	2.03
CodeLlama 2 7B	6.64	1.60	0.80	9.20	1.83	13.20	2.15	10.80	1.97	10.80	1.97	4.40	1.30	7.60	1.68	1.20	0.69	1.60	0.80	6.00	1.51	4.00	1.24
LLaMAX Llama 2 7B	3.62	3.60	1.18	3.60	1.18	7.20	1.64	3.60	1.18	5.20	1.41	2.80	1.05	2.80	1.05	0.80	0.56	0.80	0.56	2.00	0.89	4.80	1.35
LLaMAX Llama 2 7B Alpaca	6.35	3.20	1.12	4.80	1.35	15.20	2.28	9.60	1.87	5.20	1.41	4.40	1.30	3.60	1.18	4.80	1.35	0.40	0.40	4.80	1.35	6.80	1.60
MaLA-500 Llama 2 10B v1	0.73	0.00	0.00	0.40	0.40	0.40	0.40	0.40	0.40	1.60	0.80	1.60	0.80	2.40	0.97	0.40	0.40	0.00	0.00	0.00	0.00	0.80	0.56
MaLA-500 Llama 2 10B v2	0.73	0.00	0.00	0.40	0.40	0.40	0.40	0.40	0.40	1.60	0.80	1.60	0.80	2.40	0.97	0.40	0.40	0.00	0.00	0.00	0.00	0.80	0.56
YaYi Llama 2 7B	7.22	2.80	1.05	8.40	1.76	16.80	2.37	12.40	2.09	10.40	1.93	4.80	1.35	7.20	1.64	2.00	0.89	1.20	0.69	3.20	1.12	12.40	2.09
TowerBase Llama 2 7B	6.16	3.60	1.18	8.00	1.72	11.20	2.00	8.40	1.76	7.60	1.68	3.60	1.18	8.40	1.76	1.20	0.69	0.80	0.56	2.80	1.05	9.20	1.83
TowerInstruct Llama 2 7B	8.24	1.20	0.69	12.80	2.12	18.80	2.48	15.20	2.28	12.00	2.06	2.40	0.97	13.60	2.17	1.60	0.80	1.20	0.69	3.20	1.12	10.80	1.97
EMMA-500 Llama 2 7B	18.09	8.40	1.76	22.40	2.64	37.60	3.07	25.60	2.77	21.60	2.61	8.00	1.72	22.80	2.66	21.20	2.59	2.40	0.97	11.60	2.03	16.40	2.35
Occiglot Mistral 7B v0.1	14.07	2.40	0.97	21.60	2.61	33.20	2.98	26.80	2.81	22.40	2.64	6.40	1.55	17.20	2.39	3.20	1.12	1.20	0.69	6.00	1.51	11.20	2.00
Occiglot Mistral 7B v0.1 Instruct	22.16	4.00	1.24	33.20	2.98	43.20	3.14	42.00	3.13	31.60	2.95	16.80	2.37	24.40	2.72	5.60	1.46	1.60	0.80	8.40	1.76	24.80	2.74
BLOOM 7B	2.29	2.00	0.89	1.60	0.80	4.80	1.35	2.40	0.97	2.40	0.97	0.40	0.40	4.40	1.30	1.60	0.80	2.00	0.89	2.00	0.89	2.40	0.97
BLOOMZ 7B	2.15	2.00	0.89	2.00	0.89	2.40	0.97	3.20	1.12	3.20	1.12	2.00	0.89	1.60	0.80	2.40	0.97	2.00	0.89	1.60	0.80	1.20	0.69
YaYi 7B	3.02	4.80	1.35	3.20	1.12	4.80	1.35	3.60	1.18	4.00	1.24	2.00	0.89	1.20	0.69	1.60	0.80	2.80	1.05	0.80	0.56	6.00	1.51
Aya 23 8B	24.71	6.00	1.51	40.80	3.11	48.80	3.17	41.20	3.12	40.00	3.10	28.40	2.86	37.20	3.06	5.60	1.46	2.40	0.97	8.40	1.76	6.80	1.60
Aya Expanse 8B	41.45	21.20	2.59	68.80	2.94	78.80	2.59	76.80	2.68	65.20	3.02	29.60	2.89	67.60	2.97	11.20	2.00	2.40	0.97	20.00	2.53	2.80	1.05
Gemma 7B	35.78	36.40	3.05	40.80	3.11	60.80	3.09	45.20	3.15	42.00	3.13	3.60	1.18	39.20	3.09	38.80	3.09	24.80	2.74	44.00	3.15	4.80	1.35
Gemma 2 9B	44.69	48.80	3.17	53.60	3.16	72.00	2.85	60.40	3.10	50.00	3.17	2.00	0.89	50.00	3.17	53.60	3.16	44.40	3.15	55.20	3.15	1.60	0.80
Qwen 1.5 7B	30.36	12.80	2.12	43.20	3.14	58.40	3.12	50.40	3.17	42.00	3.13	11.20	2.00	36.80	3.06	8.80	1.80	4.00	1.24	25.60	2.77	26.00	2.78
Qwen 2 7B	51.47	45.20	3.15	66.00	3.00	82.00	2.43	75.20	2.74	71.60	2.86	2.80	1.05	68.40	2.95	17.20	2.39	17.60	2.41	58.40	3.12	60.80	3.09
Qwen 2.5 7B	55.60	25.60	2.77	69.60	2.92	81.60	2.46	73.60	2.79	68.00	2.96	60.00	3.10	74.80	2.75	17.20	2.39	9.20	1.83	56.80	3.14	72.80	2.82
Marco-LLM GLO 7B	52.02	32.00	2.96	65.20	3.02	78.80	2.59	70.00	2.90	64.00	3.04	48.80	3.17	64.40	3.03	18.00	2.43	9.60	1.87	56.00	3.15	63.60	3.05
Llama 3 8B	28.13	19.20	2.50	38.40	3.08	54.00	3.16	46.00	3.16	36.00	3.04	2.00	0.89	38.00	3.08	25.20	2.75	6.40	1.55	39.20	3.09	4.00	1.24
Llama 3.1 8B	27.31	19.20	2.50	42.00	3.13	52.80	3.16	42.80	3.14	34.80	3.02	2.80	1.05	37.20	3.06	25.20	2.75	5.60	1.46	35.60	3.03	1.60	0.80
LLaMAX Llama 3 8B	19.96	12.40	2.09	20.40	2.55	24.80	2.74	25.60	2.77	18.40	2.46	18.80	2.48	20.40	2.55	23.20	2.68	3.60	1.18	30.80	2.93	19.60	2.52
LLaMAX Llama 3 8B Alpaca	17.16	14.80	2.25	18.00	2.43	26.00	2.78	22.40	2.64	24.00	2.71	13.60	2.17	22.40	2.64	12.40	2.09	7.20	1.64	11.60	2.03	13.20	2.15
EMMA-500 Llama 3 8B Mono	25.33	23.20	2.68	34.00	3.00	37.60	3.07	40.40	3.11	32.80	2.98	1.60	0.80	33.60	2.99	32.40	2.97	12.80	2.12	24.00	2.71	2.80	1.05
EMMA-500 Llama 3 8B Bi	26.29	22.00	2.63	34.40	3.01	40.40	3.11	42.80	3.14	37.20	3.06	0.80	0.56	31.60	2.95	33.60	2.99	12.00	2.06	29.20	2.88	1.20	0.69
EMMA-500 Llama 3.1 8B Mono	27.35	22.00	2.63	38.00	3.08	40.00	3.10	41.60	3.12	36.40	3.05	3.20	1.12	37.20	3.06	42.40	3.13	10.40	1.93	26.40	2.79	2.00	0.89
EMMA-500 Llama 3.1 8B Bi	25.76	24.40	2.72	38.00	3.08	35.20	3.03	38.80	3.09	35.20	3.03	1.60	0.80	33.20	2.98	35.20	3.03	8.80	1.80	25.20	2.75	1.20	0.69

Table 26: 3-shot results (Accuracy %) on MGSM in all languages by CoT prompting and flexible matching.

machine translation. Kondo et al. (2024) showed that continual pre-training and supervised fine-tuning on parallel data can enhance the accuracy in English-Japanese translation. Similarly, Fujii et al. (2024) investigated cross-lingual continual pre-training on English and Japanese and showed that the use of parallel corpora enhanced translation performance. Li et al. (2025) studied the data mixing strategies for continual pre-training using monolingual and parallel texts and code data across 30+ languages. While previous studies have explored bilingual and multilingual pre-training with parallel texts, most efforts have been limited to a small subset of languages or specific translation tasks. Unlike PolyLM and Poro, which incorporate a modest fraction of bilingual data into predominantly monolingual training, our work systematically integrates bilingual texts across 500+ languages.

F Ethics Considerations

Data Collection and Use This work involves the compilation and use of large-scale monolingual and bilingual textual corpora covering over 500 languages. All datasets used for continual pre-training are obtained from publicly available sources or established linguistic resources that permit academic research use. Given the scale and diversity of the data, we acknowledge the potential presence of sensitive or harmful content. Future work could

include filtering mechanisms or annotation efforts to improve the quality and safety of multilingual data.

Representation of Low-Resource Languages Our research aims to expand LLM coverage to underrepresented languages. However, the quantity and quality of data for low-resource languages remain uneven. This may inadvertently lead to biased model behaviors or inadequate performance for some linguistic groups.

Community Involvement This work does not directly involve community collaboration. However, we recognize the importance of inclusive research practices and welcome future partnerships with linguists, native speakers, and regional institutions to improve the quality and cultural relevance of multilingual language technologies.