MaLA: A Corpus and Data Mix for Massive Language Adaptation of Large Language Models

Shaoxiong Ji[1, 2], **Zihao Li**[3], **Jaakko Paavola**[3], **Peiqin Lin**[4, 5], **Pinzhen Chen**[6], **Dayyán O'Brien**[6], **Hengyu Luo**[3], **Hinrich Schütze**[4, 5], **Jörg Tiedemann**[3], **Barry Haddow**[6]

[1]ELLIS Institute Finland [2]University of Turku [3]University of Helsinki
[4]University of Munich [5]Munich Center for Machine Learning [6]University of Edinburgh

Emails: shaoxiong.ji@utu.fi linpq@cis.lmu.de
{zihao.li, jaakko.paavola, hengyu.luo, jorg.tiedemann}@helsinki.fi
{pinzhen.chen, dobrien4, bhaddow}@ed.ac.uk

Abstract

In this work, we present the **MaLA** (**Ma**ssively multilingual **L**anguage **A**daptation) suite, a comprehensive collection of open-access resources designed to advance language technology across 546 languages. The primary contribution is the **MaLA corpus**, a 74-billion-token dataset specifically engineered for the continual pre-training of large language models (LLMs). To ensure high utility and robustness for various tasks, we introduce a **diverse data mix** that balances underrepresented languages with curated high-resource data. This mix includes: (1) structured scientific literature; (2) literary archives; (3) multilingual instruction sets; and (4) programming data.

We document the corpus provenance and sampling strategies—upsampling low-resource and downsampling high-resource languages—to reduce forgetting (particularly in code generation) while expanding language capacity. Leveraging this resource, we release **EMMA-500**, a Llama 2-based model optimized for cross-lingual transfer and language adaptability. Beyond the model weights, we provide the full suite of scripts, processing pipelines, and model generations to support reproducibility and further experimentation in multilingual indexing and retrieval. We release the MaLA suite, including the MaLA corpus, EMMA-500 model weights, scripts, and model generations, under open licenses to provide the community with the testbeds and tools necessary to evaluate and improve global information access.

Homepage: www.olaresearch.org/MaLA
Data: huggingface.co/collections/MaLA-LM/mala-corpus
Model: huggingface.co/MaLA-LM/emma-500-llama2-7b
Evaluation: github.com/MaLA-LM/emma-500

1 Introduction

Fueled by Transformer architectures (Vaswani et al., 2017) and massive multilingual corpora like ROOTS (Laurençon et al., 2022), HPLT (de Gibert et al., 2024), and FineWeb 2 (Penedo et al., 2025), multilingual large language models (LLMs) such as XLM-R (Conneau et al., 2020) and mT5 (Xue et al., 2021) have evolved rapidly. By scaling data and parameters, these models achieve strong task performance and successfully leverage cross-lingual transfer from high-resource languages to their lower-resource counterparts. However, severe data imbalances persist. While extensive corpora exist for languages like English and Spanish, data for languages such as Xhosa and Inuktitut remains scarce and fragmented.

Consequently, models trained on these skewed distributions inevitably overfit to high-resource languages, leaving low-resource populations critically underserved.

Recent studies adopt continual learning to enhance the language coverage of large language models on low-resource languages (K et al., 2026; Li et al., 2026). For example, Glot500 (Imani et al., 2023) and MaLA-500 (Lin et al., 2024) use continual pre-training and vocabulary extension using XLM-R and LLaMA, respectively, on the Glot500-c corpus covering 534 languages. xLLMs-100 (Lai et al., 2024) proceeds to multilingual instruction fine-tuning to improve the multilingual performance of LLaMA and BLOOM models on 100 languages, and Aya model (Üstün et al., 2024) applies continual training to the mT5 model (Xue et al., 2021) using their constructed instruction dataset. LLaMAX (Lu et al., 2024) pushes the envelope by focusing on translation tasks via continual pre-training of LLaMA in over 100 languages, and Gao et al. (2025) release a recent development in multilingual reasoning with parallel data in 17 languages. Li et al. (2025) systematically demonstrate the effect of data mixing in multilingual continual pretraining, particularly for low-resource languages.

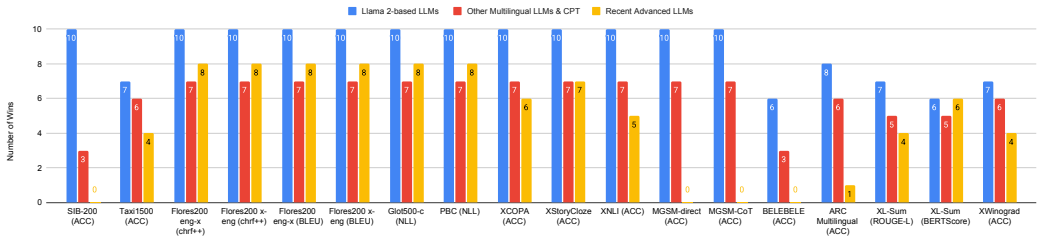


Figure 1: The number of wins, i.e., the number of times EMMA-500 achieves the best or superior performance compared to other models in the same category across various evaluation tasks and benchmarks. We compare our EMMA-500 Llama 2 7B model to decoder-only LLMs of similar parameter size, including (i) 10 Llama 2-based LLMs, (ii) 7 multilingual LLMs and CPT models, and (iii) 8 recent advanced LLMs (see Appendix C.2.2) on tasks and benchmarks in Table 6. If EMMA-500 scores higher than all compared models on a specific benchmark, it is considered a winning case for that particular evaluation. Our EMMA-500 Llama 2 model outperforms most Llama 2-based, multilingual LLMs and CPT models. Remarkably, our model achieves the best performance on Flores200, Glot500-c, and PBC among all the compared baselines.

As the field of multilingual LLMs evolves, the role of data becomes increasingly critical in enhancing the performance of the models, particularly when it comes to low-resource languages. To address this need for more and better data, we present a new release and experimental setting distinct from prior efforts like MaLA-500. While MaLA-500 (Lin et al., 2024) is a model trained specifically on the Glot500-c corpus with vocabulary expansion, this work introduces a new 939-language corpus, a 136B-token training mix, a Llama-2-based demonstration model trained without vocabulary expansion, and a significantly broader evaluation and release package. We emphasize the creation of a massively multilingual corpus that not only increases the quantity of data but also diversifies the types of texts (e.g., code, books, scientific papers, and instructions). This ensures better language adaptation and broader language coverage, thus improving the representation and performance of multilingual language models, especially for underrepresented languages, ultimately creating more inclusive and versatile language models that cater to a broader linguistic diversity. Our contribution is summarized as follows:

- We compile a massively multilingual corpus named MaLA to facilitate continual training of large language models for enhanced language adaptation across a wide range of linguistic contexts.
- We extend the MaLA corpus by integrating multiple curated datasets, creating a comprehensive and diverse data mix specifically for continual pre-training.

- We perform continual pre-training with the Llama 2 7B model[1] on the multilingual corpus with 546 languages, resulting in the new EMMA-500 model. This model is rigorously evaluated on a diverse set of tasks and benchmarks.

The MaLA Corpus Our multilingual corpus features the following characteristics:

- It contains 939 languages, (546 of which are used for training our EMMA-500 model) and 74 billion (B) whitespace delimited tokens in total.
- It has more than 300 languages with over 1 million whitespace delimited tokens and 546 languages with over 100k tokens.
- It comes with four publicly available versions: (1) noisy, (2) cleaned, (3) deduplicated, and (4) split. Rather than releasing a single opaque dump, this staged release along with our processing scripts allows practitioners to inspect, reproduce, or construct alternative custom mixes from the staged resources.
- Our augmentation to the MaLA corpus includes different types of texts such as code, books, scientific papers and instruction data, leading to a data mix with 100B+ whitespace delimited tokens.

Evaluation Results To evaluate downstream impact beyond intrinsic language modeling, we benchmark on a comprehensive suite of 9 task families across 15 benchmarks (including translation, classification, commonsense reasoning, natural language inference, summarization, math, and code generation). In a comparison with decoder-only LLMs (Appendix C.2.2) including Llama 2-based continual pre-trained models, LLMs that are designed to be multilingual, and the latest advanced LLMs, our model achieves strong results (Figure 1):

- Out of models with parameter sizes from 4.5B to 13B, our model with 7B parameters has the lowest negative log-likelihood according to an intrinsic evaluation.
- Our model remarkably improves the performance of commonsense reasoning, and machine translation over Llama 2-based models and multilingual baselines, and outperforms the latest advanced models in many cases.
- Our model improves the performance of text classification and natural language inference, outperforming all Llama 2-based models and LLMs designed to be multilingual.
- While math and machine reading comprehension (MRC) tasks are challenging for the Llama 2 7B model and other multilingual LLMs, our model remarkably enhances the Llama 2 base model. Our model yields improved performance on MRC over the base model but still produces quasi-random results similar to other multilingual baselines.
- We demonstrate that massively multilingual continued pre-training does not necessarily lead to regressions in other areas, such as code generation, if the data mix is carefully curated. Our model surpasses the Llama 2 7B base model's code generation abilities.

Novelty and Contributions The primary contribution of this work is framed around three pillars: (1) a massive, multi-stage resource (the MaLA corpus), (2) a reproducible adaptation recipe (the data mix and training methodology), and (3) an empirical demonstration model (EMMA-500). Rather than proposing new preprocessing algorithms, our focus is scale, multilingual coverage, transparent data auditing, and baseline-controlled validation. Specifically, we apply an auditable pipeline over hundreds of languages to compile a 939-language corpus (released in raw, cleaned, and deduplicated stages), construct a reproducible 136B-token training mix, and validate the resource using an architecture-matched demonstration model (EMMA-500) to isolate the impact of our data mix on adaptation. The scope and impact of this work make it a landmark contribution, particularly in expanding NLP capabilities for underrepresented languages.

[1]Choosing English-centric Llama 2 allows us to compare our model with many other models derived from it using continual pre-training, and isolate the impact of our constructed corpus and data mix on multilingual adaptation.

2 The MaLA Corpus

The MaLA corpus—MaLA standing for **Ma**ssive **L**anguage **A**daptation—is a diverse and extensive compilation of text data encompassing 939 languages sourced from a wide array of datasets. Throughout this paper, we distinguish between the **MaLA Corpus** (the curated monolingual dataset spanning 939 languages) and the **MaLA Training Mix** (the final 136B-token pre-training mix detailed in Section 3). This section introduces the efforts made in data extraction, harmonization, pre-processing, cleaning, and deduplication in order to build the corpus.

2.1 Data Integration

To develop the MaLA corpus for training our language model, we establish a processing workflow consisting of the following key steps: (1) loading and curating identified data sources, (2) extracting and harmonizing both textual and metadata from these diverse sources into a unified format—often with tailored filtering, and (3) performing deduplication and further filtering on the textual data. In step (1), source data are either organized and loaded into memory from a file tree structure or, if available, accessed through an API. In step (2), the output is designed to be `JSON` Lines (`JSONL`) files with extracted text data and other desired content. We selectively process only data annotated as training or development (validation) data, deliberately excluding test data. Our source selection for this versioned release is guided by: (1) public availability and redistributability, (2) language coverage emphasizing low- and medium-resource targets (e.g., using monoHPLT from HPLT v1.2 while excluding high-resource portions), (3) usable metadata for script/language labeling, (4) pipeline compatibility, and (5) a fixed snapshot for reproducibility (excluding concurrent scrapes like FineWeb-2).

2.1.1 Extraction and Harmonization

As mentioned, there is significant variability in the data quality of the source datasets used for compiling the corpus. Many of the datasets exhibit data quality challenges that would have adverse effects on model training if left unaddressed. These dataset-specific challenges are typically addressed during the pre-processing stage. For example, we identify one issue involving text records consisting solely of date and timestamp information in the dataset for Languages of Russia (Orekhov et al., 2016), likely resulting from a web scraper failing to differentiate between these elements and actual text. We address this by implementing a logic in the pre-processing script to detect and exclude such records.

Due to the extensive volume of data, exhaustive examination of every source for data quality issues is impractical. Instead, we address issues only as we encounter them in our data exploration and pre-processing pipeline development efforts. This approach likely leaves some data quality problems undetected, since we do not go through the data systematically.

Despite the need for customized handling of certain dataset-specific idiosyncrasies, the core logic and structure of the pre-processing workflow remain consistent across most datasets. We develop a standardized pre-processing script that can be adapted with minor adjustments to accommodate different datasets.

2.1.2 Language Code Normalization

An essential component of our pre-processing pipeline involved converting language codes to the ISO 639-3 standard. This is crucial for ensuring consistent language identification across the source datasets. We rely on the declared language of each dataset and normalize it to the ISO standard without performing additional language identification at this step. This approach helps maintain uniformity while streamlining the pre-processing workflow.

For monolingual data, we primarily use the PyPI library `iso639-lang`[2] or `langcodes`[3]. While converting language codes to ISO 639-3, we encounter several challenges. One issue is that some languages in ISO 639-3 are divided into multiple subvarieties, but our source data does not specify which subvariety is present. Our solution is to retain the original language code from the dataset, even when it does not conform to the ISO 639-3 standard.

Another issue arises when certain languages are merged into other languages in the ISO 639-3 standard. In these cases, we update the language code to reflect the merged language. Additionally, some language names or codes in the source data—referred to as "original language names" or "original language codes"—are not recognized by the conversion libraries. In some cases, the reason behind this is that the original code in fact represents language families or groups of dialects (e.g., the ISO 639-2 codes "ber" for Berber languages and "bih" for Bihari languages), rather than specific languages. If so, we then retain the original codes, despite their non-compliance with ISO 639-3. In other cases, the original language names are spelled differently from the standard recognized by the libraries. To address this mismatch, we implement a logic to detect and correct misspelled language names during pre-processing. All these "corner cases" require careful attention in the pre-processing stage to ensure correct language code identification.

2.1.3 Writing System Recognition

In addition to normalizing language codes, we also identify the script or writing system used in the text data. We use the `GlotScript` library (Kargaran et al., 2024) to recognize writing systems accordant to the ISO 15924 standard. The process begins by sampling 100 random lines from each dataset (or the full dataset if it contains fewer than 100 lines). If `GlotScript` fails to identify a script from this sample, we attempt identification using just the first line of the sample. If this still does not yield a result, we set the script as "None". It is worth noting that we choose not to classify a dataset into multiple scripts, even when code-mixing (i.e., the use of multiple scripts) is present.

If script identification is unsuccessful after the initial steps, we assume the script matched a previously detected one for that language. In cases where no previous script information exists, we refer to a mapping of languages and their default scripts provided by the Glot500 corpora collection. Through this multi-step process, we are able to determine the script for every dataset without exceptions.

During script identification, we encounter several challenges. One issue is determining an appropriate length for the text chunk used for script recognition. A chunk that is too short could lead to incorrect identification if the text contains quotes or foreign language fragments using a different writing system. Conversely, using a chunk that is too long could result in excessive resource usage, slowing down processing or even causing memory exhaustion. Another consideration is whether to assume that a single file or dataset might contain multiple scripts. Such an assumption would require identifying the script at a more granular level, such as paragraph by paragraph or even sentence by sentence. Alternatively, we could assume that each file or dataset contained only one "main" script. This assumption would allow us to identify the script from a representative sample of the text for the whole file or dataset. We adopt the latter approach, recognizing a single dominant script for each dataset. The output of this process is a label in the format language_Script, e.g., eng_Latn, where "Language" represents the ISO 639-3 language code and "Script" represents the ISO 15924 script code.

2.2 Data Cleaning

Most source data has already undergone data cleaning to different extents. Nonetheless, different cleaning processes have been adopted. We continue to clean the data to ensure consistency and accuracy for monolingual and bilingual texts.

[2] https://pypi.org/project/iso639-lang/

[3] https://pypi.org/project/langcodes

Following the pipeline used by BigScience's pipeline for ROOTS corpus (Laurençon et al., 2022), we further adopt some necessary data cleaning to filter out text samples that might have undesirable quality. We first perform document modification for monolingual texts. The first step is whitespace standardization: all types of whitespace in a document are converted into a single, consistent space character. We split documents by newline characters, tabs, and spaces, strip words, and reconstruct the documents to remove very long words. However, these two steps do not apply to languages without whitespace word delimiters like Chinese, Japanese, Korean, Thai, Lao, Burmese, etc. We also remove words containing certain patterns, e.g., "http" and ".com", which are likely to be links and page source code. We then perform document filtering, including word count filtering, character repetition filtering, word repetition filtering, special characters filtering, stop words filtering, and flag words filtering. To validate language labeling, we use the pre-trained `fastText`-based language identification (LID) model (Joulin et al., 2016b;a) as a lightweight filter only for supported languages. For the remaining low-resource languages, we rely on the normalized source metadata because automatic LID is often unsupported or unreliable here, though more advanced tools like GlotLID (Kargaran et al., 2023) represent promising alternatives for future iterations.

2.3 Data Deduplication

To remove overlaps across sources, we deduplicate each language's dataset grouped by writing system in a sequential workflow. First, we apply MinHashLSH near-deduplication (Rajaraman & Ullman, 2011) using the `text-dedup` repository (Mou et al., 2023) with 5-grams and a similarity Jaccard threshold of 0.7. Second, we perform exact deduplication on the remaining documents using MD5 hashes (Rivest, 1992). For programming code, we apply a separate pipeline using MinHash with a 0.5 threshold and shingle size 20 (details in Appendix A.2).

2.4 Key Statistics

This section presents the final MaLA corpus obtained after data sourcing, pre-processing, cleaning, and deduplication. Table 1 first shows some basic data statistics and compares them with other multilingual corpora for pre-training language models or language adaptation. Additional statistics are presented in Appendix B in the appendix. The token counts are based on white-space delimitation, though it might not be accurate for languages like Chinese, Japanese and Korean since the entire clause is counted as one token. Glot500-c (Imani et al., 2023) has 534 languages in total, in which 454 languages are directly distributed on Huggingface [4]. We also omit high-resource languages in the other three datasets, i.e., MADLAD (Kudugunta et al., 2024), CulturaX (Nguyen et al., 2023), and CC100 (Wenzek et al., 2020), as our main focus is continual pre-training for language adaptation. The final MaLA corpus consists of 939 languages, 546 of which have more than 100k tokens and are used for training our EMMA-500 model. Counting languages with more than 1 million tokens, the MaLA corpus and Glot500-c have more than 300, while MADLAD and CulturaX have 200 and 100 respectively. Compared with Glot500-c, MaLA has longer average documents (90.12 vs. 19.53 tokens) that can provide more context. However, document length is not isolated from cleaning and domain differences; the core evidence for MaLA's effectiveness is the end-to-end comparison against CPT baselines.

3 Data Mixing

Incorporating a diverse data mix—spanning various languages, domains, document lengths and styles—is crucial for continual training of large language models to enhance their versatility, generalization ability, and robustness across a wide range of tasks and domains. We construct the **MaLA Training Mix** by augmenting the MaLA corpus with

[4] https://huggingface.co/datasets/cis-lmu/Glot500

Table 1: Data statistics of the MaLA corpus and comparison to other multilingual corpora. The number of documents and tokens is in millions.

Dataset	N Lang	N Lang Counted	N Docs	N Tokens	Avg Tokens/Doc
Glot500-c	534	454	1,815	35,449	19.53
MADLAD	419	414	1,043	645,111	618.51
CulturaX	167	161	2,141	1,029,810	480.99
CC100	116	101	2,557	52,201	20.41
MaLA	939	546	824	74,255	90.12

diverse datasets to mitigate issues such as over-fitting or capacity loss during continual pre-training.

Curated Data We enhance the corpus with high-quality curated data, specifically high-resource languages in the monolingual part. We use texts from scientific papers as these provide a structured, information-dense corpus that can improve the model's ability to handle technical language and domain-specific content. They are (1) CSL (Li et al., 2022), a large-scale Chinese Scientific Literature dataset, that contains titles, abstracts, keywords and academic fields of 396,209 papers; (2) pes2o (Soldaini & Lo, 2023), a collection of full-text open-access academic papers derived from the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2020). We further add free e-books from the Gutenberg project[5] compiled by Faysse (2023). These texts enhance the range of literary styles and narrative forms, thus enhancing the model's versatility. Adding high-resource languages into the pre-training corpora also mitigates the forgetting in model training.

Instruction Data We further augmented the training corpus by incorporating instruction-based datasets, inspired by Li et al. (2023); Taylor et al. (2022); Nakamura et al. (2024). We mix two instruction data into our training corpus. They are: (1) xp3x (Crosslingual Public Pool of Prompts eXtended)[6], a multitask instruction collection in 277 languages (Muennighoff et al., 2022); (2) the Aya collection [7] that contains both human-curated and machine translated instructions in 101 languages (Singh et al., 2024). For both instruction datasets, we use their training set.

Code We additionally enrich the training corpus by sourcing code data from The Stack (Kocetkov et al., 2023). This is done following existing work that demonstrates the value of code data in improving the reasoning ability of language models (Zhang et al., 2024b; Ma et al., 2024) while also reducing forgetting on evaluated retained capabilities, particularly the base model's programming knowledge. We subsample The Stack at an effective rate of 15.2%, prioritizing high-quality source files and data science code [8]. We retain the 32 most important non-data programming languages by prevalence while also adding in all `llvm` code following prior work detailing its importance in multi-lingual code generation (Szafraniec et al., 2023; Paul et al., 2024). We also source from data-heavy formats but follow precedent (Lozhkov et al., 2024) and subsample them more aggressively. For a more detailed read on filtering heuristics, we direct the reader to Appendix A.2.

Data Mix Our final data mix for continual training is listed in Table 2. The resource categorization refers to Appendix B.1 in the appendix and `inst medium-high+` is a separate category with languages with more than 500 million but less than 1B tokens. For monolingual text, we also have a separate category for English. In continual learning, where new data is introduced to an existing model, there is a risk of forgetting previously learned skills, particularly code generation and reasoning. Although our work's primary focus is in a low-resource regime, we enhance the training corpus with a wide range of data types, including books and scientific papers in high-resource languages, code, and instruction

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

[6] https://huggingface.co/datasets/CohereForAI/xP3x

[7] https://huggingface.co/datasets/CohereForAI/aya_collection_language_split

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

Table 2: Data mix for continual training. Code and reasoning-related data are counted by Llama 2 tokenizer and others are counted as whitespace delimited; 'inst' stands for instruction and 'mono' stands for monolingual texts.

Data	Original Counts	Sample Rate	Final Counts	Percentage
inst high	42,121,055,562	0.1	4,212,105,556	3.08%
inst medium-high+	6,486,592,274	0.2	1,297,318,455	0.95%
inst medium-high	30,651,187,534	0.5	15,325,593,767	11.21%
inst medium	1,444,764,863	1.0	1,444,764,863	1.06%
inst medium-low	47,691,495	5.0	238,457,475	0.17%
inst low	3,064,796	20.0	61,295,920	0.04%
inst code/reasoning	612,208,775	1.0	612,208,775	0.45%
code	221,003,976,266	0.1	20,786,882,764	15.20%
curated (EN pes2o)	56,297,354,921	0.2	11,241,574,489	8.22%
curated (ZH CSL & wiki)	61,787,372	1.0	61,787,372	0.05%
curated (Gutenberg)	5,173,357,710	1.0	5,173,357,710	3.78%
mono high EN	3,002,029,817	0.1	300,202,982	0.22%
mono high	40,411,201,964	0.5	20,205,600,982	14.78%
mono medium-high	27,515,227,962	1.0	27,515,227,962	20.12%
mono medium	2,747,484,380	5.0	13,737,421,900	10.05%
mono medium-low	481,935,633	20.0	9,638,712,660	7.05%
mono low	97,535,696	50.0	4,876,784,800	3.57%

data in our data mix. We downsample texts in high-resource languages and upsample text in low-resource languages using different sample rates according to how resourceful the language is. Specifically, the sample rates in Table 2 reflect this balance: high-resource categories are downsampled (rates of 0.1–0.2) to prevent dominant-language bias, while low-resource categories are upsampled (rates up to 50.0) to ensure sufficient representation, with code (15.20%) and curated scientific literature (8.22%) acting as reasoning anchors. We make our data mix diverse and balanced towards different resource groups of languages in order to retain the prior knowledge of the model while learning new information, especially in medium- and low-resource languages, thus maintaining high performance across both previously seen and new languages. The final data mix has around 136B tokens. While we did not conduct exhaustive ablation studies on specific mixing ratios, the final data mix was heuristically curated to integrate these diverse text types while aggressively upsampling low-resource languages. This ensures a robust, balanced foundation that reduces forgetting on evaluated retained capabilities while driving massive language adaptation.

To link our data mix choices to experimental evidence: (1) *Tiered sampling* (NLL and translation gains in Appendices C.4 and C.5); (2) *Code replay* (retains base programming skills on Multipl-E in Appendix C.12); (3) *Multitask instructions* (boosts commonsense and NLI task adaptation in Appendices C.7 and C.8); and (4) *Vocabulary preservation* (controlled baseline comparability).

4 Model Training & Evaluation

Model Training We employ continual training using the causal language modelling objective for the decoder-only Llama model and exposing the pre-trained model to new data to develop our EMMA-500 model. We preserved the original vocabulary because CPT ablations (e.g., Lu et al., 2024) and comparisons with MaLA-500 show limited benefits from expansion. Retaining the vocabulary does not prevent effective adaptation here; we leave tokenizer fertility studies as future work. We adopt efficient training strategies combining optimization, memory management, precision handling, and distributed training techniques. Our EMMA-500 model is trained on the Leonardo supercomputer[9], occupying 256 Nvidia A100 GPUs, using the GPT-NeoX framework (Andonian et al., 2023). During training, we set a global batch size of 4096 and worked with sequences of 4096 tokens. The training process ran for 12,000 steps, resulting in a total of 200 billion Llama 2 tokens. We use the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.0001, betas of [0.9, 0.95], and an epsilon of 1e-8. We use a cosine learning rate scheduler with a warm-up of 500

[9] https://leonardo-supercomputer.cineca.eu

iterations. To reduce memory consumption, activation checkpointing is employed. Precision is managed through mixed-precision techniques, using bfloat16 for computational efficiency and maintaining FP32 for gradient accumulation.

Tasks and Benchmarks We conduct a comprehensive evaluation to validate the usability of our processed data and data mixing for massively multilingual language adaptation. We perform the intrinsic evaluation of the models' performance on next-word prediction and evaluate the model's performance on downstream tasks. Table 6 lists the datasets we used as downstream evaluation datasets in this work. For math tasks, we perform direct prompting and Chain-of-Thoughts prompting on the MGSM benchmark and evaluate the performance using both exact and flexible matches. For code generation tasks, we perform test-case-based execution-tested evaluations using the `pass@k` metric (Chen et al., 2021). We benchmark for `k` values of 1,10, and 25 using a generation pool of 50 samples per problem. For massively multilingual benchmarks with more than 100 languages, we categorize languages into five groups—high, medium-high, medium, medium-low, and low—based on their token counts in the MaLA corpus as listed in Table 5 in Appendix B.1. We compare with 25 decoder-only LLMs (Appendix C.2.2), reporting performance across all languages (including unsupported ones) to test generalizability. Instead of labor-intensive taxonomies, we group languages into five tiers directly by token count (Table 5).

Results Table 7 presents the overall performance across seven diverse tasks, ranging from text classification to reasoning and comprehension. EMMA-500 demonstrates consistently strong multilingual capabilities across both intrinsic and downstream evaluations. By adapting Llama 2, we establish a controlled comparison against other CPT adaptations (e.g., LLaMAX and MaLA-500). EMMA-500 consistently outperforms MaLA-500 v2 on NLL (Glot500-c/PBC), showing that MaLA's sequence length and curated mix compensate for not expanding the vocabulary. Furthermore, our mix bridges the gap to natively multilingual models (e.g., Gemma 2) while reducing forgetting on evaluated capabilities like code generation.

On intrinsic evaluations, EMMA-500 achieves the lowest negative log-likelihood on both the Glot500-c and PBC test sets, indicating superior multilingual text modeling (details in Appendix C.4). In topic classification tasks (Appendix C.6), it surpasses all Llama 2-based models on SIB-200 and delivers competitive results on Taxi1500. For commonsense reasoning (details in Appendix C.7), EMMA-500 outperforms Llama 2 models and performs on par with, or better than, the more recent Llama 3 series. In the natural language inference task (details in Appendix C.8), it ranks second overall, trailing only Llama 3.1. On the MGSM benchmark for multilingual math problem solving (details in Appendix C.10), EMMA-500 shows marked improvements over Llama 2 7B in both direct and chain-of-thought prompting, closing the gap with Llama 3-level performance. Although its performance on the BELEBELE reading comprehension task (details in Appendix C.11) is moderate, it still advances beyond the Llama 2 baselines. In multilingual summarization (details in Appendix C.9), EMMA-500 delivers results comparable to other recent models, though the performance margin is less distinct. Table 3 presents the results on the FLORES-200 benchmark and demonstrates that our EMMA-500 model delivers strong multilingual translation capabilities. In the X-to-English (X-Eng) translation direction, EMMA-500 achieves the highest average performance compared with all baselines. In the English-to-X (Eng-X) direction, EMMA-500 outperforms all compared models, including translation-specialized ones like Tower and LLaMAX. In summary, EMMA-500 achieves substantial improvements over Llama 2-based models, performs competitively with more recent architectures across several benchmarks, and consistently demonstrates robust generalization and cross-lingual reasoning capabilities. The findings highlight the value of our combination of compiling the MaLA corpus, mixing diverse data, and large-scale continual pretraining in promoting multi-task language understanding and generation in a massively multilingual scenario.

Qualitative Analysis and Practical Implications Synthesizing these evaluation results leads to several key qualitative insights:

- EMMA-500's gains in low-resource NLL and FLORES translation show that targeted upsampling effectively bootstraps underrepresented languages.
- In architecture-matched comparisons (against LLaMAX and MaLA-500), EMMA-500 consistently achieves superior performance.
- Replaying code in the mix reduces forgetting on evaluated capabilities like Multipl-E, whereas other CPT baselines collapse.
- Though newer models (e.g., Llama 3.1 and Gemma 2) excel in high-resource settings, this supports using MaLA to adapt future backbones.

Practical Implications for Practitioners Based on our adaptation recipe, we offer several recommendations for adapting future foundation backbones:

1. Use our staged releases (raw to split) to inspect and selectively remix data rather than treating it as a black box.
2. Apply resource-tier sampling rates to prevent low-resource language starvation.
3. Include replay-style base text, multi-task instructions, and source code to preserve reasoning and programming skills.
4. Use the MaLA corpus as a resource to adapt stronger foundation backbones.
5. Evaluate both broad-scale multilingual generalization and targeted supported-language subsets.

Further analyses on all tasks are available in Appendices C and D.

Table 3: Results on FLORES-200 including English-to-X (Eng-X) and X-to-English (X-Eng). EMMA-500 Llama 2 7B (Our) has better performance than all baselines.

Model	Eng-X chrF / BLEU	X-Eng chrF / BLEU	Model	Eng-X chrF / BLEU	X-Eng chrF / BLEU
Llama 2 7B	15.13 / 4.62	30.32 / 12.93	mGPT	12.56 / 2.59	20.69 / 5.29
Llama 2 7B Chat	16.95 / 4.95	31.72 / 12.28	mGPT 13B	14.57 / 3.88	24.58 / 7.42
CodeLlama 2 7B	14.94 / 4.27	28.57 / 10.82	YaYi 7B	13.50 / 4.37	21.36 / 4.82
LLaMAX Llama 2 7B	7.42 / 0.80	13.66 / 1.99	Aya 23 8B	16.15 / 6.46	32.36 / 13.87
LLaMAX Llama 2 7B Alpaca	28.35 / 12.51	42.27 / 22.29	Aya Expanse 8B	23.89 / 6.88	36.86 / 13.12
MaLA-500 Llama 2 10B v1	6.08 / 0.60	13.60 / 2.29	Llama 3 8B	24.08 / 9.93	43.72 / 23.78
MaLA-500 Llama 2 10B v2	6.38 / 0.54	15.44 / 2.87	Llama 3.1 8B	24.69 / 10.11	44.10 / 24.19
YaYi Llama 2 7B	14.87 / 4.41	31.38 / 12.98	Gemma 7B	23.05 / 9.05	43.68 / 23.79
TowerBase Llama 2 7B	16.03 / 4.83	31.47 / 13.74	Gemma 2 9B	26.48 / 12.09	38.87 / 23.15
TowerInstruct Llama 2 7B	15.64 / 3.23	25.43 / 4.81	Qwen 1.5 7B	17.77 / 5.87	35.87 / 15.58
Occiglot Mistral 7B v0.1	16.10 / 4.32	31.13 / 13.12	Qwen 2 7B	17.17 / 5.56	37.61 / 17.39
Occiglot Mistral 7B v0.1 Instruct	15.80 / 3.99	31.65 / 11.61	Qwen 2.5 7B	17.49 / 5.72	38.89 / 18.95
BLOOM 7B	11.80 / 2.81	27.84 / 9.57	Marco-LLM GLO 7B	23.34 / 9.27	44.57 / 25.17
BLOOMZ 7B	16.10 / 7.44	34.74 / 20.22	**EMMA-500 (Ours)**	**33.25 / 15.58**	**45.78 / 25.37**

5 Conclusion

In this work, we have presented the **MaLA suite**, a foundational resource designed to bridge the gap in multilingual information access across 546 languages. By meticulously curating a 136-billion-token data mix—integrating scientific literature, literary prose, multi-task instructions, and programming code—we have provided the community with a framework for massively multilingual language adaptation. Our methodology emphasizes a balanced sampling strategy that preserves high-resource knowledge while significantly expanding capacity in low-resource regimes. The resulting **EMMA-500** model serves as a high-utility artifact that demonstrates the effectiveness of our corpus in enhancing cross-lingual transfer and task generalization. By releasing the MaLA corpus, model weights, and the full preprocessing and training pipelines, we aim to lower the barrier to entry for researchers working on underrepresented languages.

Acknowledgments

The work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070350 and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546], and funding from UTTER's Financial Support for Third Parties under the EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631). The work has also received funding from the Digital Europe Programme under grant agreement No 101195233 (OpenEuroLLM). The authors wish to acknowledge CSC - IT Center for Science, Finland, the Leonardo and LUMI supercomputers, owned by the EuroHPC Joint Undertaking, for providing computational resources. We thank Indraneil Paul for his contributions to preparing the code data and evaluating our model on code generation tasks. Shaoxiong Ji gratefully acknowledges the support of Foundation PS through the PS Fellowship.

References

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. Parallel corpora for bi-lingual English-Ethiopian languages statistical machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3102–3111, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1262.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pp. 1–10, Kyiv, Ukraine (Virtual), April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wanlp-1.1.

Kathrein Abu Kwaik, Motaz Saad, Stergios Chatzikyriakidis, and Simon Dobnik. Shami: A corpus of Levantine Arabic dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1576.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pp. 61–75, Virtual, August 2021. Association for Machine Translation in the Americas. URL https://aclanthology.org/2021.mtsummit-research.6.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3053–3070, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL https://aclanthology.org/2022.naacl-main.223.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *CoRR*, abs/2309.07445, 2023. doi: 10.48550/arXiv.2309.07445. URL https://doi.org/10.48550/arXiv.2309.07445.

Rodrigo Agerri, Xavier Gómez Guinovart, German Rigau, and Miguel Anxo Solla Portela. Developing new linguistic resources and tools for the Galician language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1367.

Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed. DART: A large dataset of dialectal Arabic tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1579.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*, 2024. URL https://arxiv.org/abs/2402.17733.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcovid19-2.5. URL https://aclanthology.org/2020.nlpcovid19-2.5.

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, Phil Wang, and Samuel Weinbach. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch, 9 2023. URL https://www.github.com/eleutherai/gpt-neox.

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

Mikko Aulamo, Nikolay Bogoychev, Shaoxiong Ji, Graeme Nail, Gema Ramírez-Sánchez, Jörg Tiedemann, Jelmer Van Der Linde, and Jaume Zaragoza. Hplt: High performance language technologies. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pp. 517–518, 2023. URL https://aclanthology.org/2023.eamt-1.61/.

Niyati Bafna. Empirical models for an indic language continuum, 2022.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The BELEBELE benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*, 2023. URL https://aclanthology.org/2024.acl-long.44/.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting*

of the Association for Computational Linguistics, pp. 4555–4567, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.417. URL https://aclanthology.org/2020.acl-main.417.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubesic, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. Macocu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In Helena Moniz, Lieve Macken, Andrew Rufener, Loïc Barrault, Marta R. Costa-jussà, Christophe Declercq, Maarit Koponen, Ellie Kemp, Spyridon Pilos, Mikel L. Forcada, Carolina Scarton, Joachim Van den Bogaert, Joke Daems, Arda Tezcan, Bram Vanroy, and Margot Fonteyne (eds.), *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, EAMT 2022, Ghent, Belgium, June 1-3, 2022*, pp. 301–302. European Association for Machine Translation, 2022. URL https://aclanthology.org/2022.eamt-1.41.

José Camacho-Collados, Claudio Delli Bovi, Alessandro Raganato, and Roberto Navigli. A large-scale multilingual disambiguation of glosses. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 1701–1708, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1269.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q Feldman, et al. Multipl-e: A scalable and extensible approach to benchmarking neural code generation. *arXiv preprint arXiv:2208.08227*, 2022.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. When is multilinguality a curse? Language modeling for 250 high- and low-resource languages. *arXiv preprint*, 2023. URL https://arxiv.org/abs/2311.09205.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

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

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2018.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020.

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

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, et al. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of LREC-COLING*, 2024.

Aleksei Dorkin, Taido Purason, Emil Kalbaliyev, Hele-Andra Kuulmets, Marii Ojastu, Mark Fišel, Tanel Alumäe, Eleri Aedmaa, Krister Kruusmaa, and Kairit Sirts. Estllm: Enhancing estonian capabilities in multilingual llms via continued pretraining and post-training, 2026. URL https://arxiv.org/abs/2603.02041.

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

Jonathan Dunn. Mapping languages: the corpus of global language use. *Lang. Resour. Evaluation*, 54(4):999–1018, 2020. doi: 10.1007/s10579-020-09489-2. URL https://doi.org/10.1007/s10579-020-09489-2.

EdTeKLA. Indigenous languages corpora, 2022. URL https://github.com/EdTeKLA/IndigenousLanguages_Corpora.

Mahmoud El-Haj. Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1318–1326, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.165.

Mahmoud El-Haj, Paul Rayson, and Mariam Aboelezz. Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1573.

Manuel Faysse. Dataset card for "project gutenberg", 2023. URL https://huggingface.co/datasets/manu/project_gutenberg.

Fitsum Gaim, Wonsuk Yang, and Jong C. Park. Tlmd: Tigrinya language modeling dataset (1.0.0), 2021. URL https://doi.org/10.5281/zenodo.5139094. Dataset.

Changjiang Gao, Zixian Huang, Jingyang Gong, Shujian Huang, Lei Li, and Fei Yuan. LLaMAX2: Your Translation-Enhanced Model also Performs Well in Reasoning, October 2025.

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

Javier García Gilabert, Carlos Escolano, Aleix Sant Savall, Francesca De Luca Fornaciari, Audrey Mash, Xixian Liao, and Maite Melero. Investigating the translation capabilities of large language models trained on parallel data only, 2024. URL https://arxiv.org/abs/2406.09140.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 759–765, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. Experiments on a Guarani corpus of news and social media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pp. 153–158, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.americasnlp-1.16. URL https://aclanthology.org/2021.americasnlp-1.16.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. Can we use word embeddings for enhancing Guarani-Spanish machine translation? In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 127–132, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.computel-1.16. URL https://aclanthology.org/2022.computel-1.16.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 306–316, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.37. URL https://aclanthology.org/2021.acl-demo.37.

Grupo de Inteligencia Artificial PUCP. Monolingual and parallel corpora of peruvian languages, n.d. URL https://github.com/iapucp/multilingual-data-peru.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4693–4703, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.413. URL https://aclanthology.org/2021.findings-acl.413.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models, 2024. URL https://arxiv.org/abs/2403.08763.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. Masakhaner: Named entity recognition for african languages. *arXiv e-prints*, pp. arXiv–2103, 2021.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André F. T. Martins, François Yvon, and Hinrich Schütze. Glot500: Scaling multilingual corpora and language models to 500 languages. *arXiv preprint*, 2023. URL https://aclanthology.org/2023.acl-long.61/.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016a.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016b.

Santosh Srinath K, Mudit Somani, Varun Reddy Padala, Prajna Devi Upadhyay, and Abhijit Das. Continual-learning for modelling low-resource languages from large language models, 2026. URL https://arxiv.org/abs/2601.05874.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4948–4961, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.445. URL https://aclanthology.org/2020.findings-emnlp.445.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. GlotLID: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=dl4e3EBz5j.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. Glotscript: A resource and tool for low resource writing system identification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 7774–7784, 2024.

Najoung Kim, Sebastian Schuster, and Shubham Toshniwal. Code pretraining improves entity tracking abilities of language models. *CoRR*, abs/2405.21068, 2024. doi: 10.48550/ARXIV.2405.21068. URL https://doi.org/10.48550/arXiv.2405.21068.

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

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. The stack: 3 TB of permissively licensed source code. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=pxpbTdUEpD.

Minato Kondo, Takehito Utsuro, and Masaaki Nagata. Enhancing translation accuracy of large language models through continual pre-training on parallel data, 2024. URL https://arxiv.org/abs/2407.03145.

Fajri Koto and Ikhwan Koto. Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pp. 138–148, Hanoi, Vietnam, October 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.paclic-1.17.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 2024.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1548.

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

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 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*, pp. 318–327, 2023.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 8186–8213, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.488.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 8608–8621. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.emnlp-main.590.

Hector Levesque, Ernest Davis, and Leora Morgenstern. The Winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

Haolin Li, Haipeng Zhang, Mang Li, Yaohua Wang, Lijie Wen, Yu Zhang, and Biqing Huang. Toward robust multilingual adaptation of llms for low-resource languages, 2026. URL https://arxiv.org/abs/2510.14466.

Shenggui Li, Hongxin Liu, Zhengda Bian, Jiarui Fang, Haichen Huang, Yuliang Liu, Boxiang Wang, and Yang You. Colossal-ai: A unified deep learning system for large-scale parallel training. In *Proceedings of the 52nd International Conference on Parallel Processing*, pp. 766–775, 2023.

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

Zihao Li, Shaoxiong Ji, Hengyu Luo, and Jörg Tiedemann. Rethinking multilingual continual pretraining: Data mixing for adapting llms across languages and resources. In *Conference on Language Modeling (COLM)*, 2025.

Evenki Life. Evenki life newspaper, 2014. URL https://drive.google.com/file/d/1he2q6RncA_NKHPIJjSzlkK-2qgEFTiCG/view.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

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

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

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://www.aclweb.org/anthology/2020.acl-main.447.

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

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. LLaMAX: Scaling linguistic horizons of LLM by enhancing translation capabilities beyond 100 languages. *arXiv preprint arXiv:2407.05975*, 2024.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *CoRR*, abs/2308.08747, 2023. doi: 10.48550/ARXIV.2308.08747. URL https://doi.org/10.48550/arXiv.2308.08747.

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

Yingwei Ma, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. At which training stage does code data help llms reasoning? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=KIPJKST4gw.

Martin Majliš. W2C – web to corpus – corpora, 2011. URL http://hdl.handle.net/11858/00-097C-0000-0022-6133-9. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Masakhane. Lacuna project, 2023. URL https://github.com/masakhane-io/lacuna_pos_ner.

Thomas Mayer and Michael Cysouw. Creating a massively parallel bible corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pp. 3158–3163. European Language Resources Association (ELRA), 2014. URL http://www.lrec-conf.org/proceedings/lrec2014/summaries/220.html.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Bekhzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5876–5890, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.475. URL https://aclanthology.org/2021.emnlp-main.475.

Steven Moran, Christian Bentz, Ximena Gutierrez-Vasques, Olga Pelloni, and Tanja Samardzic. TeDDi sample: Text data diversity sample for language comparison and multilingual NLP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1150–1158, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.123.

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 3603–3609, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.443.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, 2017.

Chenghao Mou, Chris Ha, Kenneth Enevoldsen, and Peiyuan Liu. Chenghaomou/text-dedup: Reference snapshot, 2023. URL https://doi.org/10.5281/zenodo.8364980.

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

Jonathan Mukiibi, Andrew Katumba, Joyce Nakatumba-Nabende, Ali Hussein, and Joshua Meyer. The makerere radio speech corpus: A luganda radio corpus for automatic speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1945–1954, 2022.

Taishi Nakamura, Mayank Mishra, Simone Tedeschi, Yekun Chai, Jason T Stillerman, Felix Friedrich, Prateek Yadav, Tanmay Laud, Vu Minh Chien, Terry Yue Zhuo, et al. Aurora-m: The first open source multilingual language model red-teamed according to the us executive order. *arXiv preprint arXiv:2404.00399*, 2024.

Toshiaki Nakazawa, Hideki Nakayama, Chenchen Ding, Raj Dabre, Shohei Higashiyama, Hideya Mino, Isao Goto, Win Pa Pa, Anoop Kunchukuttan, Shantipriya Parida, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. Overview of the 8th workshop on Asian translation. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pp. 1–45, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.wat-1.1. URL https://aclanthology.org/2021.wat-1.1.

Toshiaki Nakazawa, Hideya Mino, Isao Goto, Raj Dabre, Shohei Higashiyama, Shantipriya Parida, Anoop Kunchukuttan, Makoto Morishita, Ondřej Bojar, Chenhui Chu, Akiko Eriguchi, Kaori Abe, Yusuke Oda, and Sadao Kurohashi. Overview of the 9th workshop on Asian translation. In *Proceedings of the 9th Workshop on Asian Translation*, pp. 1–36, Gyeongju, Republic of Korea, October 2022. International Conference on Computational Linguistics. URL https://aclanthology.org/2022.wat-1.1.

Graham Neubig. The Kyoto free translation task. http://www.phontron.com/kftt, 2011.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages, 2023.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.mrl-1.11.

B. Orekhov, I. Krylova, I. Popov, E. Stepanova, and L. Zaydelman. Russian minority languages on the web: Descriptive statistics. In *Proceedings of the Annual International Conference "Dialogue" (2016) – Computational Linguistics and Intellectual Technologies*, pp. 498–508, M.: RSUH, 2016. URL http://web-corpora.net/wsgi3/minorlangs/download.

OSCAR. Oscar (open super-large crawled aggregated corpus) 2301, 2023. URL https://huggingface.co/datasets/oscar-corpus/OSCAR-2301.

Chester Palen-Michel, June Kim, and Constantine Lignos. Multilingual open text release 1: Public domain news in 44 languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2080–2089, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.224.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Indraneil Paul, Goran Glavas, and Iryna Gurevych. Ircoder: Intermediate representations make language models robust multilingual code generators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15023–15041. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.802.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language, 2025. URL https://arxiv.org/abs/2506.20920.

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

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina (eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6319.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Anand Rajaraman and Jeffrey D Ullman. *Mining of massive datasets*. Autoedicion, 2011.

Ronald Rivest. The md5 message-digest algorithm, 1992.

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

Roberts Rozis and Raivis Skadiņš. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pp. 263–265, Gothenburg, Sweden, May 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-0235.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. AraBench: Benchmarking dialectal Arabic-English machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5094–5107, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.447. URL https://aclanthology.org/2020.coling-main.447.

Ixak Sarasua, Ander Corral, and Xabier Saralegi. DIPLomA: Efficient adaptation of instructed LLMs to low-resource languages via post-training delta merging. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24898–24912, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.1355. URL https://aclanthology.org/2025.findings-emnlp.1355/.

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

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1351–1361, Online, April 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.115. URL https://aclanthology.org/2021.eacl-main.115.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6490–6500, Online, 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.507. URL https://aclanthology.org/2021.acl-long.507.

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. Language models are multilingual chain-of-thought reasoners, 2022.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. mGPT: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*, 2022.

Anil Kumar Singh. Named entity recognition for south and south East Asian languages: Taking stock. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008. URL https://aclanthology.org/I08-5003.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11521–11567, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.620.

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

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache, 2019.

Marc Szafraniec, Baptiste Rozière, Hugh Leather, Patrick Labatut, François Charton, and Gabriel Synnaeve. Code translation with compiler representations. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=XomEU3eNeSQ.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

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

Atula Tejaswi, Nilesh Gupta, and Eunsol Choi. Exploring design choices for building language-specific llms, 2024. URL https://arxiv.org/abs/2406.14670.

Huu Nguyen Thuat Nguyen and Thien Nguyen. Culturay: A large cleaned multilingual dataset of 75 languages, 2024.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

Jörg Tiedemann. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 1174–1182, Online, November 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.wmt-1.139.

Alexey Tikhonov and Max Ryabinin. It's All in the Heads: Using Attention Heads as a Baseline for Cross-Lingual Transfer in Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.

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

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.494.

Wikimedia Foundation. Wikimedia downloads, n.d. URL https://dumps.wikimedia.org.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, X. Li, Zhi Yuan Lim, S. Soleman, R. Mahendra, Pascale Fung, Syafri Bahar, and A. Purwarianti. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020.

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

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 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*, pp. 483–498, 2021.

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

Ke Yang, Jiateng Liu, John Wu, Chaoqi Yang, Yi R. Fung, Sha Li, Zixuan Huang, Xu Cao, Xingyao Wang, Yiquan Wang, Heng Ji, and Chengxiang Zhai. If LLM is the wizard, then code is the wand: A survey on how code empowers large language models to serve as intelligent agents. *CoRR*, abs/2401.00812, 2024b. doi: 10.48550/ARXIV.2401.00812. URL https://doi.org/10.48550/arXiv.2401.00812.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. Introducing QuBERT: A large monolingual corpus and BERT model for Southern Quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pp. 1–13, Hybrid, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deeplo-1.1. URL https://aclanthology.org/2022.deeplo-1.1.

Chen Zhang, Mingxu Tao, Quzhe Huang, Jiuheng Lin, Zhibin Chen, and Yansong Feng. Mc^2: Towards transparent and culturally-aware nlp for minority languages in china. *arXiv preprint arXiv:2311.08348*, 2024a.

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

Xinlu Zhang, Zhiyu Zoey Chen, Xi Ye, Xianjun Yang, Lichang Chen, William Yang Wang, and Linda Ruth Petzold. Unveiling the impact of coding data instruction fine-tuning on large language models reasoning. *CoRR*, abs/2405.20535, 2024b. doi: 10.48550/ARXIV.2405.20535. URL https://doi.org/10.48550/arXiv.2405.20535.

Contents

A Data Sources

A.1 Monolingual Data

A.1.1 List of Data Sources

The MaLA corpus harvests a wide range of datasets in multiple domains. Table 4 in the appendix lists the corpora and collections we used as monolingual data sources. The major ones include AfriBERTa (Ogueji et al., 2021), Bloom library (Leong et al., 2022), CC100 (Conneau et al., 2020; Wenzek et al., 2020) CulturaX (Nguyen et al., 2023), CulturaY (Thuat Nguyen & Nguyen, 2024), the Curse of Multilinguality (Chang et al., 2023), Evenki Life (Life, 2014), Glot500 (Imani et al., 2023), GlotSparse (Kargaran et al., 2023), monoHPLT of HPLT v1.2 (de Gibert et al., 2024) from the HPLT project (Aulamo et al., 2023), Indigenous Languages Corpora (EdTeKLA, 2022), Indo4B (Wilie et al., 2020), Lacuna Project (Masakhane, 2023), Languages of Russia (Orekhov et al., 2016), MADLAD-400 (Kudugunta et al., 2024), Makerere Radio Speech Corpus (Mukiibi et al., 2022), masakhaner1.0 (Ifeoluwa Adelani et al., 2021), MC2 (Zhang et al., 2024a), mC4 (Raffel et al., 2020), multilingual-data-peru (Grupo de Inteligencia Artificial PUCP, n.d.), OSCAR 2301 (OSCAR, 2023) from the OSCAR project [10], Tatoeba challenge monolingual collection (Tiedemann, 2020), The Leipzig Corpora (Goldhahn et al., 2012), Tigrinya Language Modeling (Gaim et al., 2021), Wikipedia 20231101 (Wikimedia Foundation, n.d.) and Wikisource 20231201 (Wikimedia Foundation, n.d.). We exclude high-resource languages in CulturaX, HPLT, MADLAD-400, CC100, mC4, and OSCAR 2301. We exclude Gahuza and Pidgin in the AfriBERTa dataset. We filter out texts that mainly contain a date or timestamp in the Languages of Russia dataset. For Glot500-c, we filter out texts, which may come from train or test tests from datasets for machine translation, such as Flores200, Tatoeba, and mtdata. Despite the translation data being split into source and target languages in the Glot500-c corpus, we decide to filter them to avoid potential data leakage, especially since we use Flores200 as the evaluation benchmark.

Table 4 lists the corpora and collections we use as monolingual data sources in this work. Monolingual data sources simply contain text data in a single language.

Metadata For the monolingual data sources, we define the contents of the `JSONL` output file from the pre-processing workflow to consist of the fields `url`, `text`, `collection`, `source`, and `original_code`. The contents of these fields are as follows. The field `text` contains the language data in a granularity specific to the given corpus. If the granularity was sentence-level, then we could expect the sentences in the corpus to generally be independent of each other, while only parts of the sentences—such as phrases, clauses, and words—exhibit serial dependence. If the granularity was paragraph-level, then we could expect the sentences within paragraphs to have serial dependence, while paragraphs to largely be independent of each other. The field `url` contains a URL indicating the web address from which the text data has been extracted, if available. The field `collection` contains the name of the collection, i.e., a corpus or a set of corpora, which the text is extracted from, whereas the field `source` contains the name of a more specific part of the collection, such as the name of an individual corpus or a file in the collection the text was extracted from. Lastly, the field

[10] https://oscar-project.org/

`original_code` contains the language code of the text data as it is designated in the data source, e.g., in the directory structure, the filenames, or the data object returned by an API call.

Table 4: Datasets used as monolingual source data.

Name	Languages	Domains	URL	Year
AfriBERTa (Ogueji et al., 2021)	10	news	https://huggingface.co/datasets/castorini/afriberta-corpus	2021
Bloom library (Leong et al., 2022)	363	religious, books	https://huggingface.co/datasets/sil-ai/bloom-lm	2022
CC100 (Conneau et al., 2020)	100	web	https://huggingface.co/datasets/cc100	2020
CulturaX (Nguyen et al., 2023)	167	web	https://huggingface.co/datasets/uonlp/CulturaX	2023
CulturaY (Thuat Nguyen & Nguyen, 2024)	75	web	https://huggingface.co/datasets/ontocord/CulturaY	2024
curse-of-multilinguality (Chang et al., 2023)	200	misc	https://github.com/tylerachang/curse-of-multilinguality	2023
Evenki Life (Life, 2014)	1	newspapers	https://drive.google.com/file/d/1he2q6RncA_NKHPIJjSzlkK-2qgEFTiCG/view	2014
Glot500 (Imani et al., 2023)	511	misc	https://huggingface.co/datasets/cis-lmu/Glot500	2023
GlotSparse (Kargaran et al., 2023)	10	news	https://huggingface.co/datasets/cis-lmu/GlotSparse	2023
HPLT v1.2 (de Gibert et al., 2024)	75	web	https://hplt-project.org/datasets/v1.2	2024
Indigenous Languages Corpora (EdTeKLA, 2022)	1	UNK	https://github.com/EdTeKLA/IndigenousLanguages_Corpora	2022
Indo4B (Wilie et al., 2020)	1	misc	https://github.com/IndoNLP/indonlu?tab=readme-ov-file	2020
Lacuna Project (Masakhane, 2023)	20	UNK	https://github.com/masakhane-io/lacuna_pos_ner	2023
Languages of Russia (Orekhov et al., 2016)	46	social media, web	http://web-corpora.net/wsgi3/minorlangs/download	UNK
Madlad-400 (Kudugunta et al., 2024)	419	web	https://huggingface.co/datasets/allenai/MADLAD-400	2023
Makerere Radio Speech Corpus (Mukiibi et al., 2022)	1	transcription	https://zenodo.org/records/5855017	2022
masakhane-ner1.0 (Ifeoluwa Adelani et al., 2021)	12	UNK	https://github.com/masakhane-io/masakhane-ner	2021
MC2 (Zhang et al., 2024a)	4	web	https://huggingface.co/datasets/pkupie/mc2_corpus	2024
mC4 (Raffel et al., 2020)	101	web	https://huggingface.co/datasets/allenai/c4	2020
multilingual-data-peru (Grupo de Inteligencia Artificial PUCP, n.d.)	4	UNK	https://github.com/iapucp/multilingual-data-peru	2020
OSCAR 2301 (OSCAR, 2023)	152	web	https://huggingface.co/datasets/oscar-corpus/OSCAR-2301	2023
The Leipzig Corpora (Goldhahn et al., 2012)	136	newspapers, wikipedia	https://huggingface.co/datasets/imvladikon/leipzig_corpora_collection	2012
Tigrinya Language Modeling (Gaim et al., 2021)	1	news, blogs, books	https://zenodo.org/records/5139094	2021
Wikipedia 20231101 (Wikimedia Foundation, n.d.)	323	wikipedia	https://huggingface.co/datasets/wikimedia/wikipedia	2023
Wikisource 20231201 (Wikimedia Foundation, n.d.)	73	books	https://huggingface.co/datasets/wikimedia/wikisource	2023
Tatoeba challenge monolingual (Tiedemann, 2020)	280	wikimedia	https://github.com/Helsinki-NLP/Tatoeba-Challenge/blob/master/data/MonolingualData-v2020-07-28.md	2020

Glot500-c uses the following datasets: AI4Bharat,[11] AIFORTHAI-LotusCorpus,[12] Add (El-Haj et al., 2018), AfriBERTa (Ogueji et al., 2021), AfroMAFT (Adelani et al., 2022; Xue et al., 2021), Anuvaad,[13] AraBench (Sajjad et al., 2020), AUTSHUMATO,[14] Bloom (Leong et al., 2022), CC100 (Conneau et al., 2020), CCNet (Wenzek et al., 2020), CMU_Haitian_Creole,[15] CORP.NCHLT,[16] Clarin,[17] DART (Alsarsour et al., 2018), Earthlings (Dunn, 2020), FFR,[18] Flores200 (Costa-jussà et al., 2022), GiossaMedia (Góngora et al., 2022; 2021), Glosses (Camacho-Collados et al., 2016), Habibi (El-Haj, 2020), HinDialect (Bafna, 2022), HornMT,[19] IITB (Kunchukuttan et al., 2018), IndicNLP (Nakazawa et al., 2021), Indiccorp (Kakwani et al., 2020), isiZulu,[20] JParaCrawl (Morishita et al., 2020), KinyaSMT,[21] LeipzigData (Goldhahn et al., 2012), Lindat,[22] Lingala_Song_Lyrics,[23] Lyrics,[24] MC4 (Raffel et al., 2020), MTData (Gowda et al., 2021), MaCoCu (Bañón et al., 2022), Makerere MT Corpus,[25] Masakhane community,[26] Mburisano_Covid,[27] Menyo20K (Adelani et al., 2021), Minangkabau corpora (Koto & Koto, 2020), MoT (Palen-Michel et al., 2022), NLLB_seed (Costa-jussà et al., 2022), Nart/abkhaz,[28] OPUS (Tiedemann, 2012), OSCAR (Suárez et al., 2019), ParaCrawl (Bañón et al., 2020), Parallel Corpora for Ethiopian Languages (Abate et al., 2018), Phontron (Neubig, 2011), QADI (Abdelali et al., 2021), Quechua-IIC (Zevallos et al., 2022), SLI_GalWeb.1.0 (Agerri et al., 2018), Shami (Abu Kwaik

[11] https://ai4bharat.org/
[12] https://github.com/korakot/corpus/releases/download/v1.0/AIFORTHAI-LotusCorpus.zip
[13] https://github.com/project-anuvaad/anuvaad-parallel-corpus
[14] https://autshumato.sourceforge.net/
[15] http://www.speech.cs.cmu.edu/haitian/text/
[16] https://repo.sadilar.org/handle/20.500.12185/7
[17] https://www.clarin.si/
[18] https://github.com/bonaventuredossou/ffr-v1/tree/master/FFR-Dataset
[19] https://github.com/asmelashteka/HornMT
[20] https://zenodo.org/record/5035171
[21] https://github.com/pniyongabo/kinyarwandaSMT
[22] https://lindat.cz/faq-repository
[23] https://github.com/espoirMur/songs_lyrics_webscrap
[24] https://lyricstranslate.com/
[25] https://zenodo.org/record/5089560
[26] https://github.com/masakhane-io/masakhane-community
[27] https://repo.sadilar.org/handle/20.500.12185/536
[28] https://huggingface.co/datasets/Nart/abkhaz_text

et al., 2018), Stanford NLP,[29] StatMT,[30] TICO (Anastasopoulos et al., 2020), TIL (Mirzakhalov et al., 2021), Tatoeba,[31] TeDDi (Moran et al., 2022), Tilde (Rozis & Skadiņš, 2017), W2C (Majliš, 2011), WAT (Nakazawa et al., 2022), WikiMatrix (Schwenk et al., 2021a), Wikipedia,[32] Workshop on NER for South and South East Asian Languages (Singh, 2008), XLSum (Hasan et al., 2021). We filter out Flores200 when processing the Glot500-c dataset. Glot500-c includes texts in languages, i.e., Azerbaijani, Gujarati, Igbo, Oromo, Rundi, Tigrinya and Yoruba, from the XLSum dataset.

A.2 Code

All the programming language splits are filtered using the following conditions:

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

Subsequently, an aggressive MinHash deduplication pipeline with a threshold of 0.5 and a shingle size of 20 is applied. Finally, the resultant language splits are then capped at 5 million samples each.

B Additional Statistics of MaLA Corpus

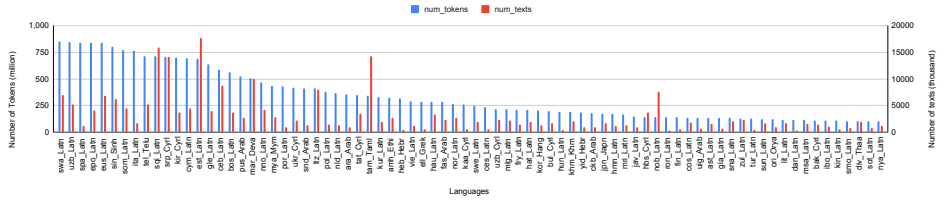
B.1 Supported Languages

Table 5 shows the languages codes of MaLA corpus, 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. Figure 2 shows the number of texts and tokens in different resource groups.

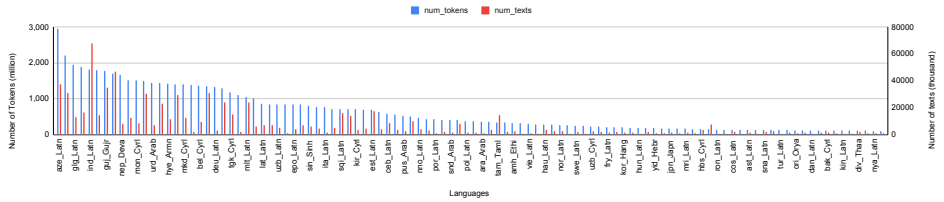
Rather than relying on predefined external taxonomies, we categorize languages into five resource tiers (High to Low) based strictly on their token frequencies within the curated MaLA corpus. This data-driven approach directly reflects the real-world availability of text data within our pre-training pipeline. To ensure precision and interoperability, we identify languages using standard ISO 639-3 codes paired with their respective writing systems. This systematic labeling effectively resolves linguistic ambiguities, particularly for macro-languages that encompass multiple dialects or utilize multiple scripts.

[29] https://nlp.stanford.edu/
[30] https://statmt.org/
[31] https://tatoeba.org/en/
[32] https://huggingface.co/datasets/wikipedia

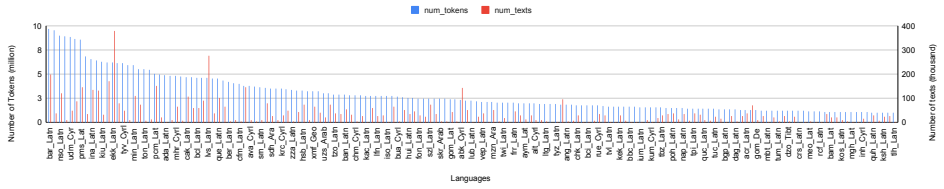
(a) High



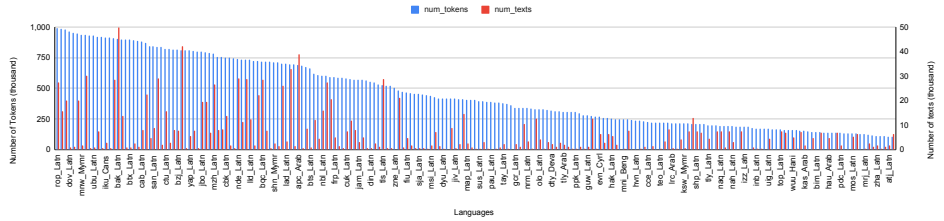
(b) Medium-high



(c) Medium



(d) Medium-low



(e) Low

Figure 2: The number of texts and tokens of MaLA corpus in different resource groups.

B.2 Data Analysis

We examine the Unicode block distribution of each language, which counts the percentage of tokens falling into the Unicode block of each language. This aims to check whether language code conversion and writing system recognition are reasonably good. Figure 3 shows the Unicode block distribution. The result aligns with our observation of the presence of code-mixing, but in general, the majority of languages have tokens falling in their own Unicode blocks.

We also check the data source distribution as shown in Figure 4 to see where the texts in the MaLA corpus come from. The main source is common crawl, e.g., CC 2018, CC, OSCAR, and CC 20220801. A large number of documents come from Earthlings which comes from

Glot500-c (Imani et al., 2023). For web-crawled data with a URL in the original metadata of corpora like CulturaX (Nguyen et al., 2023) and HPLT (de Gibert et al., 2024), we extract the domain. Thus, the final corpus has many sources with a small portion.

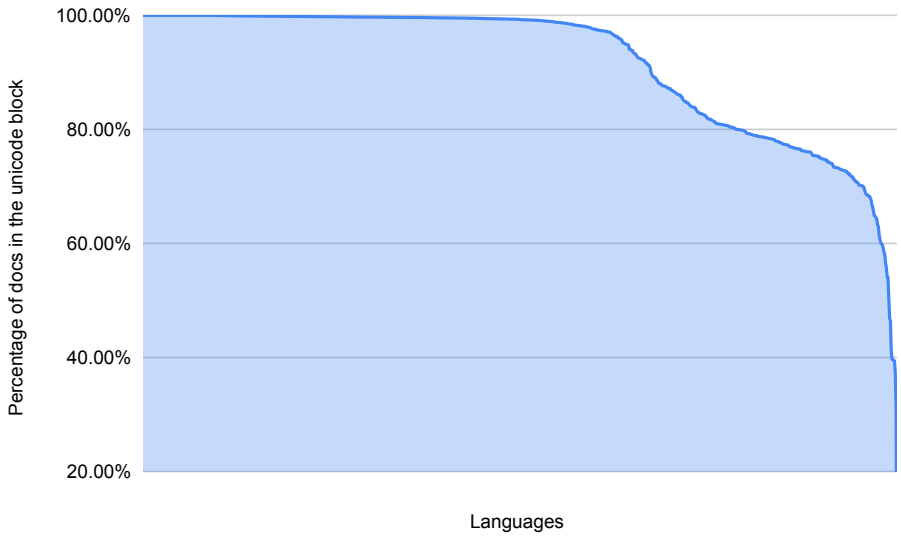


Figure 3: Unicode block distribution that measures the percentage of token counts falling into the Unicode block of each language

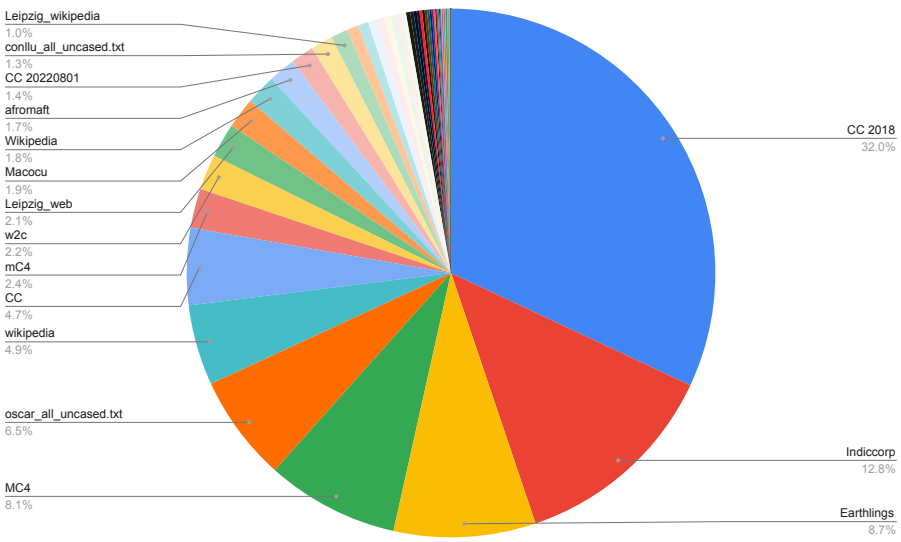


Figure 4: Data source distribution of MaLA corpus calculated by the number of documents

C Evaluation

C.1 Tasks, Benchmarks, and Baselines

Tasks and Benchmarks We conduct a comprehensive evaluation to validate the usability of our processed data and data mixing for massively multilingual language adaptation. We perform the intrinsic evaluation of the models' performance on next-word prediction and evaluate the model's performance on downstream tasks. Table 6 lists the datasets we

used as downstream evaluation datasets in this work. For math tasks, we perform direct prompting and Chain-of-Thoughts prompting on the MGSM benchmark and evaluate the performance using both exact and flexible matches. For code generation tasks, we perform test-case-based execution-tested evaluations using the `pass@k` metric (Chen et al., 2021). We benchmark for `k` values of 1,10, and 25 using a generation pool of 50 samples per problem. For massively multilingual benchmarks with more than 100 languages, we categorize languages into five groups—high, medium-high, medium, medium-low, and low—based on their token counts in the MaLA corpus as listed in Table 5 in Appendix B.1.

C.2 Prompts

C.2.1 Machine Translation

We use the prompt below for machine translation.

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

C.2.2 Text Classification

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
```

Evaluation Software We use the Language Model Evaluation Harness (`lm-evaluation-harness`) framework (Gao et al., 2023) for benchmarking test sets that are already ingested in the framework. For other benchmarks, we use in-house developed evaluation scripts and other open-source implementations. In text classification tasks such as SIB-200 and Taxi-1500, the evaluation protocol involves calculating the probability of the next output token for each candidate category. These probabilities are then sorted in descending order, with the category having the highest probability being selected as the model's prediction. This process is implemented using the Transformers (Wolf et al., 2019) library. In tasks like Machine Translation and Open-Ended Generation, the vLLM library (Kwon et al., 2023) is used to accelerate inference, providing significant speed improvements. Similarly, for code generation tasks, we use a VLLM-enabled evaluation harness package (`vllm-code-harness`)[33] for the execution-tested evaluations.

Baselines We compare our model with three groups of decoder-only models. They are (1) Llama 2 models (Touvron et al., 2023) and continual pre-trained models based on Llama 2, such as CodeLlama (Roziere et al., 2023), MaLA-500 (Lin et al., 2024), LLaMAX (Lu et al., 2024), Tower (Alves et al., 2024), and YaYi[34]; (2) other LLMs and continual pre-trained LLMs designed to be massively multilingual, including BLOOM (Scao et al., 2022), mGPT (Shliazhko et al., 2022), and Occiglot[35]; and (3) recent LLMs with superior English capabilities like Llama 3 (Dubey et al., 2024), Llama 3.1[36], Qwen 2 (Yang et al., 2024a), Aya

[33] https://github.com/iNeil77/vllm-code-harness

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

[35] https://huggingface.co/occiglot/occiglot-7b-eu5

[36] https://llama.meta.com/docs/model-cards-and-prompt-formats/llama3_1

23 (Aryabumi et al., 2024) and Gemma 2 (Team et al., 2024). There are also some other LLMs such as OpenAI's API models [37] and xLLMs-100 (Lai et al., 2024). However, they do not release the model weights or they limit access to them through commercial API, so we did not include them. The MADLAD model (Kudugunta et al., 2024) that uses the decoder-only T5 architecture is not supported by inference engines such as the HuggingFace transformers (Wolf et al., 2019). We do not compare them in this work.

C.3 Overall Results

Table 7 presents the overall results across seven distinct tasks, with performance metrics averaged by language to provide a comprehensive view of cross-lingual capabilities. The evaluation suite covers a broad spectrum of natural language processing domains, including commonsense and causal reasoning via XCOPA (XC), XStoryCloze (XSC), and XWinoGrad (XWG), alongside massive-scale reading comprehension through BELEBELE (BELE) and scientific reasoning using the multilingual AI2 Reasoning Challenge (ARC). For generative capabilities, the XL-Sum summarization task is reported using both semantic (BERTScore) and lexical (ROUGE-L) metrics. Finally, mathematical reasoning is evaluated using the Multilingual Grade School Math (MGSM) benchmark under two distinct settings: Direct prompting (Dir.) and Chain-of-Thought prompting (CoT), while Glot refers specifically to the Glot500-c corpus.

C.4 Intrinsic Evaluation

For intrinsic evaluation, we compute the negative log-likelihood (NLL) of the test text given by the tested LLMs instead of using length-normalized perplexity due to different text tokenization schemes across models. We concatenate the test set as a single input and run a sliding-window approach.[38] To make a comparison across models, we use the Glot500-c test set, which covers 534 evaluated languages. We also test on the Parallel Bible Corpus (PBC) which could yield NLL more comparable across languages. Table 8 and Table 9 show the intrinsic evaluation results on Glot500-c test and PBC, respectively. As shown, our model attains lower NLL compared to all other models on both test sets and languages with different resource availability. This suggests that the test sets are more similar to the underlying training data of our model and it can be interpreted that EMMA-500 has learned to compress massive multilingual text more efficiently.

C.5 Machine Translation

FLORES-200 is an evaluation benchmark for translation tasks with 204 language pairs involving English and thus 408 translation directions, with a particular focus on low-resource languages. We assess all language models by adopting a 3-shot evaluation approach with the prompt in Appendix C.2.1. The performance is measured by BLEU (Papineni et al., 2002) and chrF++ (Popović, 2015) implemented in `sacrebleu` (Post, 2018). The BLEU score is calculated with the `flores200` tokenizer applied to the texts and chrF++ uses word order 2. The choice of `flores200` tokenization ensures that languages that do not have a whitespace delimiter can be evaluated at the (sub-)word level. For reproducibility, we attach the BLEU and chrF++ signatures.[39]

Table 10 presents the average X-to-English (X-Eng) translation results.[40] Our EMMA-500 model outperforms all other models on average. We achieve the best performance across all language settings, except for high-resource languages where our model slightly lags

[37] https://platform.openai.com/docs/models

[38] https://huggingface.co/docs/transformers/en/perplexity

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

[40] We mark BLOOMZ with a † because it has used FLORES in its instruction tuning data; we mark MaLA-500 with a ‡ because it has used FLORES in its training data but with source and target sides split. Besides, as a remark, Tower, LLaMAX, and our EMMA-500 have intentionally used parallel data (not FLORES) in the training stage.

behind Llama 3/3.1, Gemma 7B, and LLaMAX 7B Alpaca. In the English-to-X (Eng-X) translation direction, as shown in Table 11, the advantage of EMMA-500 is even more pronounced. We outperform all other models even in high-resource languages, and the advantage becomes more significant in lower-resource languages. Overall, we note that our model outperforms Tower models which are explicitly adjusted to perform translation tasks in high-resource languages. Further, the much larger margin between EMMA-500 and other models in Eng-X compared with X-Eng indicates that our EMMA-500 model is particularly good at generating non-English texts.

C.6 Text Classification

SIB-200 (Adelani et al., 2023) and Taxi1500 (Ma et al., 2023) are two prominent topic classification datasets. SIB-200 encompasses seven categories: science/technology, travel, politics, sports, health, entertainment, and geography. Taxi1500 spans 1507 languages, involving six classes: Recommendation, Faith, Description, Sin, Grace, and Violence.

We use 3-shot prompting with prompts in Appendix C.2.2, drawing demonstrations from the development set and testing models on the test split. The outcomes are tabulated in Table 12 for SIB-200 and Taxi1500. For SIB-200, our EMMA-500 model outperforms all Llama2-based models, with particularly notable gains in languages with medium or fewer resources—seeing an average improvement of 47.5%. Taxi-1500 could be a more challenging task since it is in the religious domain, but our model still surpasses all Llama2-based models except for MaLA-500. However, despite these improvements in both classification tasks, our models lag behind the latest models such as Llama3 and 3.1, especially in high-resource languages.

C.7 Commonsense Reasoning

We assess the models' commonsense reasoning ability using three multilingual benchmarks. XCOPA (Ponti et al., 2020) is a dataset covering 11 languages that focuses on causal commonsense reasoning across multiple languages. XStoryCloze (Lin et al., 2022) is derived from the English StoryCloze dataset (Mostafazadeh et al., 2017) and translated into 10 non-English languages, testing commonsense reasoning within a story. In this task, the system must choose the correct ending for a four-sentence narrative. XWinograd (Tikhonov & Ryabinin, 2021) is a multilingual collection of Winograd Schemas (Levesque et al., 2012) available in six languages, aimed at evaluating cross-lingual commonsense reasoning. We perform zero-shot evaluations. Accuracy is used as the evaluation metric. We categorize languages into different resource groups based on language availability, accessibility, and possible corpus size. For XCOPA, we have three groups, i.e., high-resource (Italian, Turkish, Vietnamese, and Chinese), medium-resource (Swahili due to its regional influence, Estonian, Haitian Creole, Indonesian, Thai, and Tamil), and low-resource languages (Cusco-Collao Quechua).[41] For XStoryCloze, the resource group is high-resource (Arabic, English, Spanish, Russian, and Chinese), medium (Hindi, Indonesian, Swahili, and Telugu), and low (Basque and Burmese). For XWinograd, there is only one group for high-resource languages. Table 13 shows the evaluation results of zero-shot commonsense reasoning.

Compared with Llama 2-based models on XCOPA, our model improves the average performance by a large margin—up to a 5% increase when compared with the best-performing TowerInstruct based on Llama 2 7B.[42] Our model also outperforms all the multilingual LLMs. Recent LLMs such as Gemma and Llama 3 have stronger performance than Llama 2 models, and Gemma 2 9B performs the best. However, our model achieves better performance than Qwen, Llama 3, and 3.1. We gain similar results on XStoryCloze, outperforming all the models except Gemma 2 9B. As for XWinograd, a multilingual benchmark with high-resource only, our model achieves improved performance than Llama 2, despite not being as good as Tower models, which target high-resource languages. However, our

[41] Note that there is no perfect categorization for language resource groups.

[42] The biggest improvements are on medium-resource languages. However, we note that the resource categorization is not perfect.

model is comparable to other multilingual LLMs. For low-resource languages, our model outperforms all the compared LLMs except LLaMAX Llama 2 7B Alpaca on XCOPA, where we achieve the same accuracy as it.

C.8 Natural Language Inference

We evaluate on the XNLI (Cross-lingual Natural Language Inference) benchmark (Conneau et al., 2018) where sentence pairs in different languages need to be classified as entailment, contradiction, or neutral. We categorize the languages in XNLI into 3 groups, i.e., high-resource (German, English, Spanish, French, Russian, and Chinese), medium-resource (Arabic, Bulgarian, Greek, Hindi, Turkish, and Vietnamese), and low-resource (Swahili, Thai, and Urdu).[43] Table 14 shows the aggregated accuracy. Our model outperforms most baselines including Llama 2-based models and multilingual LLMs. We achieve the second-best average accuracy, slightly behind the Llama 3.1 8B model. On the low-resource end, we perform the second-best, slightly behind the Gemma 2 9B model.

C.9 Summarization

XL-Sum (Hasan et al., 2021) is a large-scale multilingual abstractive summarization dataset that covers 44 languages. To evaluate the quality of the generated summaries, we use ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2019) as evaluation metrics. ROUGE-L measures the longest common subsequence (LCS) between the reference and generated summaries. Recall and precision are calculated by dividing the LCS length by the reference length and the generated summary length, respectively. An F-score is then used to combine these two aspects into a single metric. BERTScore computes the semantic similarity between summaries by leveraging contextual embeddings from pre-trained language models. Specifically, we use the `bert-base-multilingual-cased`[44] model for BERTScore to accommodate multiple languages.

The evaluation results are presented in Table 15, we observe that our model, EMMA-500, performs comparably to other Llama2-based models like TowerInstruct Llama 2 7B. Comparing with recent advances, our EMMA-500 gets slightly better performance than Llama 3 and Gemma. While EMMA-500 shows strength in certain language categories, it does not significantly outperform these models overall. This suggests that, although EMMA-500 is effective in multilingual summarization tasks, there is room for improvement to achieve more consistent performance across all languages. It is also important to note that this test set includes only 44 languages, limiting the evaluation of EMMA-500's capabilities in low-resource languages.

C.10 Math

We evaluate the ability of LLMs to solve grade-school math problems across multiple languages on the MGSM (Multilingual Grade School Math) benchmark (Shi et al., 2022). MGSM is an extension of the GSM8K (Grade School Math 8K) dataset (Cobbe et al., 2021) by translating 250 of the original GSM8K problems into ten languages. We also split these ten languages into three groups, i.e., high-resource (Spanish, French, German, Russian, Chinese, and Japanese), medium-resource (Thai, Swahili, and Bengali), and low-resource (Telugu). Table 16 shows the results for 3-shot prompting with a flexible match to obtain the answers in model generation. We evaluate all the models by directly prompting the questions (denoted as direct) and the questions with answers followed by Chain-of-Thoughts prompt in the same languages as the subset being evaluated (denoted as CoT) (Shi et al., 2022). The results show that Llama 2 7B is a weak model in the MGSM math task. The base model and its variants failed in most settings. Multilingual LLMs such as BLOOM, XGLM, and YaYi are also subpar at this task. Recent LLMs like Aya, Llama 3, Qwen, and Gemma obtain reasonable performance, and Qwen series models are the best. Our model

[43] Again, the categorization is not perfect.

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

improves the Llama 2 7B model remarkably in both prompt strategies. For direct prompting, our model has an average accuracy of 0.1702, 7% higher than the Llama 2 7B chat model. For CoT-based prompting, our model increases the score of the Llama 2 7B chat model from 0.1 to 0.3 (20% higher) and slightly outperforms Llama 3 and 3.1 8B models.

C.11 Machine Reading Comprehension

BELEBELE (Bandarkar et al., 2023) is a machine reading comprehension dataset covering 122 languages including high- and low-resource languages. Each question offers four multiple-choice answers based on a short passage from the FLORES-200 dataset. This benchmark is very challenging, even the English version of it presents remarkable challenges for advanced models. Table 17 shows the zero-shot results in different resource groups. Our continual pre-training improves the Llama 2 7B base model. But Llama 2 7 B-based models mostly fail in this task and get quasi-random results. Recent advanced models like Aya expanse, Llama 3.1 and Qwen 2 get reasonable results.

We then move to a more challenging task, the ARC multilingual test, which is a machine-translated benchmark (Lai et al., 2023) from the ARC dataset (Clark et al., 2018) that contains English science exam questions for multiple grade levels. We test the five-shot performance, with results shown in Table 18. The evaluation results show a similar pattern to BELEBELE that EMMA-500 improves Llama 2 but the 7B model is not capable of this challenging task, all Llama2 7B-based models obtain close to random results, and recent advances like Llama 3, Gemma, and Qwen get much better results.

C.12 Code Generation

We conduct code generation evaluations on the Multipl-E (Cassano et al., 2022) benchmark in the interest of measuring the effects of massively multilingual continual pre-training on a model's code generation utility and detecting if any catastrophic forgetting (Luo et al., 2023) has occurred on this front. Importantly, this also has implications for a model's reasoning (Yang et al., 2024b) and entity-tracking abilities (Kim et al., 2024).

Table 19 outlines comparisons against strong, openly available multilingual models such as Qwen (Yang et al., 2024a) and Aya (Üstün et al., 2024) along with other continually pre-trained models. Our main takeaway is that by carefully curating high-quality code data as part of the data mix, it is possible to not only avoid the catastrophic forgetting that has plagued existing continually pre-trained models (MaLA-500, LLaMAX and TowerBase) but also surpass the base model itself. However, the pre-training phase is still the most important, and closing the gap with stronger base models like Qwen and Gemma remains elusive.

D Detailed Results

This section presents detailed results of the evaluation. For benchmarks consisting of multiple languages, we host the results on GitHub[45]. Those benchmarks include ARC multilingual, BELEBELE, Glot500-c test set, PBC, SIB-200, Taxi-1500, FLORES200, and XLSum. For FLORES200 and XLSum, we also release the generated texts of all compared models.

Tables 27 and 28 show 3-shot results on MGSM by direct and Chain-of-Thought prompting respectively. All the scores are obtained by flexible matching.

E Limitations

Multilingual Benchmark Multilingual language models are typically designed to serve users of different languages, which reflect the diverse cultural and linguistic backgrounds of their speakers. However, many available multilingual benchmarks, including some used

[45] https://github.com/MaLA-LM/emma-500/tree/main/evaluation_results

in this study, have been created through human or machine translation. They tend to feature topics and knowledge primarily from English-speaking communities and carry imperfections due to translation, which affects the integrity of LLM evaluation (Chen et al., 2024). We highlight this discrepancy and call for collaborative efforts to develop large-scale natively-created multilingual test sets that more accurately assess the breadth of languages and cultures these models aim to cater to.

Human Evaluation While human evaluation is valuable, it comes with challenges such as subjectivity, inconsistency, and high costs, especially in a massively multilingual setting. Recruiting linguistically skilled annotators is often impractical, particularly for low-resource languages, making large-scale human evaluation unsustainable. Even assessing a subset of languages remains prohibitively expensive. While our work is constrained by these limitations, we recognize the importance of human evaluation in complementing automated metrics. However, given these challenges, we choose automatic tools rather than human evaluation to provide a more scalable and consistent alternative, despite their imperfections. Furthermore, because EMMA-500 is released as a foundational base model rather than an aligned, user-facing assistant, we defer comprehensive safety and human-preference alignment testing (e.g., red-teaming) to future downstream deployment and fine-tuning efforts.

Model Performance and Evaluation Context Our EMMA-500 model serves as a baseline-controlled demonstration model to validate the MaLA corpus and training mix, rather than claiming absolute leaderboard supremacy. We deliberately selected Llama 2 7B as a controlled substrate because many existing CPT baselines share this architecture, allowing us to isolate the impact of our data mix. While recent general-purpose backbones (e.g., Llama 3/3.1, Gemma 2, and Qwen 2/2.5) remain stronger due to larger pre-training budgets, they represent a moving target. Importantly, these newer models often improve reasoning and instruction-following faster than they expand low-resource language coverage. This highlights the forward-looking role of the MaLA corpus: while EMMA-500 is a temporary demonstration model, MaLA is a long-lived resource and adaptation recipe that remains complementary to the reasoning and multimodal landscape. Even in a reasoning-centric or multimodal era, systems still heavily rely on multilingual text for instructions, retrieval, tool use, and user interactions. MaLA provides an auditable text dataset and a clean testbed to adapt future reasoning and multimodal backbones to underrepresented languages. In future work, we plan to apply our adaptation recipe to newer state-of-the-art backbones, investigate multilingual instruction tuning, and incorporate specialized multilingual reasoning datasets to address the model's current limitations in complex mathematical reasoning and machine reading comprehension.

Real-world Usage This research focuses on expanding multilingual corpora and enhancing a model's multilingual capabilities through continual pre-training. However, it is not yet ready for real-world deployment. The model has not undergone human preference alignment or red-teaming for safety and robustness. While our work makes progress in improving multilingual NLP, further refinement, including alignment with human preferences and thorough adversarial testing, would be necessary before practical deployment.

F Related Work

Multilingual LLMs Multilingual large language models (LLMs) have made significant progress in processing and understanding multiple languages within a unified framework. Models like mT5 (Xue et al., 2021) and XGLM (Lin et al., 2022) leverage both monolingual and multilingual datasets to perform tasks such as translation and text summarization across a wide spectrum of languages. However, the predominant focus on English has led to disparities in performance, particularly for low-resource languages. Recent work on multilingual LLMs, such as BLOOM (Scao et al., 2022), has shown that adapting these English-centric models through vocabulary extension based on multilingual corpora and continual pre-training (CPT) can improve performance across languages, especially low-resource ones. These models highlight the importance of efficient tokenization and adap-

tation, which can bridge the performance gap between high-resource and low-resource languages. Moreover, in massively multilingual models, imbalanced data distributions and poor data quality often exacerbate the "curse of multilinguality" (Conneau et al., 2020), where model capacity is spread too thin or dominated by high-resource languages. Overcoming this requires transparent filtering, normalization, and curated replay to balance representation and preserve base reasoning abilities during CPT.

Multilingual Corpora The availability and use of multilingual corpora play a crucial role in training multilingual LLMs. CC100 Corpus (Conneau et al., 2020), launched in 2020, encompasses hundreds of billions of tokens and over 100 languages. Further, CC100-XL Corpus (Lin et al., 2022), created for the training of XGLM, extends across 68 Common Crawl Snapshots and 134 languages, aiming to balance language presentation and improve performance in few-shot and zero-shot tasks. The ROOTS Corpus (Laurençon et al., 2022), released in July 2022, supports BLOOM with approximately 341 billion tokens across 46 natural languages. It emphasizes underrepresented languages such as Swahili and Catalan, drawing from diverse sources including web crawls, books and academic publications. Besides, Occiglot Fineweb [46], which began to be released in early 2024, consists of around 230 million documents in 10 European languages. It combines curated and cleaned web data to support efficient training for both high- and low-resource European languages. Additionally, recent efforts parallel corpus construction from web crawls, such as ParaCrawl (Bañón et al., 2020) and CCMatrix (Schwenk et al., 2021b), have contributed to large-scale multilingual training too.

Continual Pre-training Continual pre-training has become a popular strategy to adapt LLMs to new languages and domains without retraining from scratch. The process involves updating model parameters incrementally using a relatively small amount of new data from target languages. Recent studies, such as those by Tejaswi et al. (2024), have demonstrated the effectiveness of CPT in improving the adaptability of LLMs in low-resource language settings. Techniques such as vocabulary augmentation and targeted domain-specific pre-training have been shown to significantly improve both efficiency and performance, particularly when large multilingual models are adapted to new languages. Despite its benefits, CPT can lead to catastrophic forgetting, where models lose previously learned information. To address the potential degraded performance issue, Ibrahim et al. (2024) presents a simple yet effective approach to continual pre-train models, demonstrating that with a combination of learning rate re-warming, re-decaying, and replay of previous data, it is possible to match the performance of fully re-trained models. Also, recent studies have delved into the effectiveness of continual pre-training with parallel data. As highlighted in the study by Gilabert et al. (2024), the use of a Catalan-centric parallel dataset has enabled the training of models good at trasnalting in various directions. Besides, the research by Kondo et al. (2024), proposed a two-phase continual training approach with parallel data. In the first phase, a pre-trained LLM is continually pre-trained on parallel data, followed by a second phase of supervised fine-tuning with a small amount of high-quality parallel data. Their experiments with a 3.8B-parameter model across various data formats revealed that alternating between source and target sentences during continual pre-training is crucial for enhancing translation accuracy in the corresponding direction. Sarasua et al. (2025) introduced DIPLomA, establishing that continual pre-training on monolingual low-resource data, followed by weight-merging post-training deltas. Furthermore, recent empirical studies on models like EstLLM (Dorkin et al., 2026) confirm that strategically constructed data mixtures during continued pre-training can drastically enhance specific low-resource capabilities.

[46] https://occiglot.eu/posts/occiglot-fineweb/

Table 5: Languages by resource groups

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 6: Evaluation statistics. Sample/Lang: average number of test samples per language; N Lang: number of languages covered; NLL: negative log-likelihood; ACC: accuracy.

Tasks	Dataset	Metric	Samples/Lang	N Lang	Domain
Intrinsic Evaluation (Appendix C.4)	Glot500-c test (Imani et al., 2023)	NLL	1000	534	Misc
	PBC (Mayer & Cysouw, 2014)	NLL	500	370	Bible
Text Classification (Appendix C.6)	SIB200 (Adelani et al., 2023)	ACC	204	205	Misc
	Taxi1500 (Ma et al., 2023)	ACC	111	1507	Bible
Commonsense Reasoning (Appendix C.7)	XCOPA (Ponti et al., 2020)	ACC	600	11	Misc
	XStoryCloze (Lin et al., 2022)	ACC	1870	11	Misc
	XWinograd (Tikhonov & Ryabinin, 2021)	ACC	741.5	6	Misc
Natural Language Inference (Appendix C.8)	XNLI (Conneau et al., 2018)	ACC	2490	15	Misc
Machine Translation (Appendix C.5)	FLORES-200 (Costa-jussà et al., 2022)	BLEU, chrF++	1012	204	Misc
Summarization (Appendix C.9)	XL-Sum (Hasan et al., 2021)	ROUGE-L, BERTScore	2537	44	News
Math (Appendix C.10)	MGSM direct (Shi et al., 2022)	ACC	250	10	Misc
	MGSM CoT (Shi et al., 2022)	ACC	250	10	Misc
Machine Comprehension (Appendix C.11)	BELEBELE (Bandarkar et al., 2023)	ACC	900	122	Misc
	ARC multilingual (Lai et al., 2023)	ACC	1170	31	Misc
Code Generation (Appendix C.12)	Multipl-E (Cassano et al., 2022)	Pass@k	164	7	Misc

Table 7: Overall results across 7 tasks, with different evaluation metrics averaged by language. The abbreviations are as follows. XC: XCOPA, XSC: XStoryCloze, XWG: XWinoGrad, BERT: XL-Sum with BERTScore as the evaluation metrics, R-L: XL-Sum evaluated by ROUGE-L, BELE: BELEBELE, ARC: the multilingual AI2 Reasoning Challenge, Dir.: MGSM by direct prompting, CoT: MGSM by CoT prompting, Glot: Glot500-c.

Model	Classification↑		Commonsense↑			NLI↑	Summarization↑		Comprehension↑		Math↑		Intrinsic↓	
	SIB	Taxi	XC	XSC	XWG	XNLI	BERT	R-L	BELE	ARC	Dir.	CoT	Glot	PBC
Llama 2 7B	22.41	17.54	56.67	57.55	72.47	40.19	66.52	7.11	26.27	27.56	6.69	6.36	190.58	122.10
Llama 2 7B Chat	25.58	15.44	55.85	58.41	69.45	38.58	68.44	8.84	29.05	28.02	10.22	10.91	218.87	139.14
CodeLlama 2 7B	23.35	17.03	54.69	55.68	70.92	40.19	65.74	7.15	27.38	25.23	5.93	6.64	193.96	123.98
LLaMAX Llama 2 7B	10.61	23.52	54.38	60.36	67.49	44.27	64.59	5.29	23.09	26.09	3.35	3.62	187.37	117.39
LLaMAX Llama 2 7B Alpaca	27.89	15.09	56.60	63.83	69.86	45.09	69.24	10.11	24.48	31.06	5.05	6.35	169.12	107.81
MaLA-500 Llama 2 10B v1	23.25	25.27	53.09	53.07	65.89	38.11	63.96	5.45	22.96	21.16	0.91	0.73	155.62	103.20
MaLA-500 Llama 2 10B v2	19.30	23.39	53.09	53.07	65.89	38.11	64.28	5.44	22.96	21.16	0.91	0.73	151.25	101.67
YaYi Llama 2 7B	24.57	17.73	56.71	58.42	74.50	41.28	67.21	7.98	28.32	28.40	7.09	7.22	192.98	123.57
TowerBase Llama 2 7B	19.34	17.73	56.33	57.78	74.29	39.84	67.09	7.65	26.36	27.94	6.15	6.16	192.98	123.70
TowerInstruct Llama 2 7B	20.53	17.29	57.05	59.24	74.00	40.36	68.46	8.89	27.93	30.10	7.24	8.24	199.44	127.30
Occiglot Mistral 7B v0.1	32.69	22.26	56.67	58.10	74.61	42.35	66.20	7.33	30.16	29.77	13.31	14.07	191.11	121.64
Occiglot Mistral 7B v0.1 Instruct	34.31	18.76	56.55	59.39	72.93	40.81	66.96	8.31	32.05	30.88	22.76	22.16	193.83	123.18
BLOOM 7B	17.82	14.76	56.89	59.30	70.13	41.60	64.78	6.99	24.11	23.65	2.87	2.29	202.95	129.55
BLOOMZ 7B	29.73	16.96	54.87	57.12	67.95	37.13	69.82	11.15	39.32	23.95	2.55	2.15	217.32	137.72
mGPT	27.11	10.72	55.04	54.43	59.69	40.51	59.74	3.85	23.96	20.24	1.35	1.42	340.37	225.14
mGPT 13B	33.20	17.23	56.18	56.44	63.59	41.59	62.20	4.96	24.47	21.76	1.31	1.53	282.46	180.54
YaYi 7B	35.76	16.12	56.64	60.67	69.79	39.87	69.74	12.06	37.97	24.44	2.76	3.02	226.67	143.80
Aya 23 8B	41.50	22.64	55.13	60.93	65.84	43.12	66.79	8.68	40.08	31.08	22.29	24.71	207.85	131.60
Aya Expanse 8B	57.01	18.73	56.38	64.80	71.94	45.56	68.73	10.51	46.98	36.56	43.02	41.45	206.75	130.80
Llama 3 8B	63.70	21.73	61.71	63.41	76.84	44.97	67.08	8.47	40.73	34.80	27.45	28.13	156.36	102.55
Llama 3.1 8B	61.42	20.20	61.71	63.58	75.52	45.62	66.97	8.57	45.19	34.93	28.36	27.31	154.59	101.43
Gemma 7B	58.21	13.83	63.64	65.01	77.41	42.58	64.52	6.70	43.37	38.68	38.22	35.78	692.25	460.86
Gemma 2 9B	46.25	18.05	66.33	67.67	80.07	46.74	65.45	7.38	54.49	44.15	32.95	44.69	320.81	197.41
Qwen 1.5 7B	47.95	7.29	59.44	59.85	72.59	39.47	69.13	9.58	41.83	28.93	31.56	30.36	195.52	124.02
Qwen 2 7B	54.95	21.87	60.31	61.46	76.44	42.77	69.35	10.18	49.31	33.82	48.95	51.47	188.55	120.44
Qwen 2.5 7B	53.89	17.87	61.71	62.06	77.66	43.31	69.69	10.41	54.11	35.30	53.78	55.60	184.52	117.70
Marco-LLM GLO 7B	64.15	21.99	62.45	63.87	78.99	43.99	70.41	11.46	53.95	36.34	51.85	52.02	175.01	112.55
EMMA-500 Llama 2 7B	31.27	19.82	63.11	66.38	72.80	45.14	67.20	8.58	26.75	29.53	17.02	18.09	112.20	68.11

Table 8: NLL on Glot500-c test. EMMA-500 Llama 2 7B has better average performance than all baselines.

Model	Avg	High	Medium-High	Medium	Medium-Low	Low
Llama 2 7B	190.58	146.43	176.30	205.86	210.41	196.45
Llama 2 7B Chat	218.87	173.67	204.78	239.18	240.97	223.86
CodeLlama 2 7B	193.96	146.43	180.20	210.09	212.70	200.60
LLaMAX Llama 2 7B	187.37	108.58	142.84	197.74	212.22	203.83
LLaMAX Llama 2 7B Alpaca	169.12	94.84	123.90	173.54	193.83	187.34
MaLA-500 Llama 2 10B v1	155.62	127.51	153.93	173.02	166.01	158.12
MaLA-500 Llama 2 10B v2	151.25	123.82	147.69	167.46	161.20	155.13
YaYi Llama 2 7B	192.98	149.32	179.10	208.96	212.80	198.77
TowerBase Llama 2 7B	192.98	150.41	180.19	209.12	212.18	198.89
TowerInstruct Llama 2 7B	199.44	157.33	186.42	216.92	218.82	204.93
Occiglot Mistral 7B v0.1	191.11	159.48	185.53	209.26	207.67	194.07
Occiglot Mistral 7B v0.1 Instruct	193.83	162.31	188.20	212.41	210.81	196.58
BLOOM 7B	202.95	160.33	195.01	216.15	220.46	206.89
BLOOMZ 7B	217.32	178.12	210.99	235.53	235.60	220.65
mGPT	340.37	311.14	337.29	388.97	367.45	343.14
mGPT 13B	282.46	254.78	276.03	327.15	317.14	277.13
Yayi 7B	226.67	181.48	217.34	243.42	246.58	231.65
Aya 23 8B	207.85	175.18	201.33	227.29	225.01	210.90
Aya expanse 8B	206.75	162.19	191.52	229.32	225.36	211.71
Llama 3 8B	156.36	102.78	129.36	153.11	167.26	173.56
Llama 3.1 8B	154.59	101.23	127.57	150.87	164.81	172.05
Gemma 7B	692.25	583.39	721.77	817.40	729.61	689.60
Gemma 2 9B	320.81	348.26	351.08	380.49	338.00	303.62
Qwen 2 7B	188.55	132.47	171.50	200.26	210.21	196.78
Qwen 1.5 7B	195.52	141.37	181.41	212.10	217.16	202.46
EMMA-500 Llama 2 7B	112.20	81.78	100.89	122.53	99.28	109.25

Table 9: NLL on PBC. EMMA-500 Llama 2 7B has better average performance than all baselines.

Model	Avg	High	Medium-High	Medium	Medium-Low	Low
Llama 2 7B	122.10	91.30	99.41	112.31	133.08	135.34
Llama 2 7B Chat	139.14	108.40	115.78	129.79	149.54	152.82
CodeLlama 2 7B	123.98	93.27	101.83	113.47	134.65	137.52
LLaMAX Llama 2 7B	117.39	69.41	79.06	103.74	131.90	138.03
LLaMAX Llama 2 7B Alpaca	107.81	60.05	69.36	93.39	122.44	128.77
MaLA-500 Llama 2 10B v1	103.20	94.04	98.60	100.53	105.04	107.65
MaLA-500 Llama 2 10B v2	101.67	92.42	96.30	98.88	103.34	106.62
YaYi Llama 2 7B	123.57	93.22	100.95	114.01	134.44	136.76
TowerBase Llama 2 7B	123.70	93.64	101.44	114.72	134.41	136.77
TowerInstruct Llama 2 7B	127.30	98.21	105.17	118.65	137.53	140.28
Occiglot Mistral 7B v0.1	121.64	95.15	101.86	114.37	131.49	132.93
Occiglot Mistral 7B v0.1 Instruct	123.18	96.86	103.41	115.88	132.98	134.48
BLOOM 7B	129.55	96.62	111.22	115.33	138.19	143.03
BLOOMZ 7B	137.72	107.03	119.89	125.85	145.27	150.95
mGPT	225.14	211.35	203.84	219.75	229.91	239.82
mGPT 13B	180.54	164.98	160.25	179.81	186.99	191.31
YaYi 7B	143.80	108.79	123.20	130.60	152.37	158.73
Aya 23 8B	131.60	102.16	109.05	121.83	141.64	145.46
Aya expanse 8B	130.80	96.67	104.73	122.42	142.47	145.57
Llama 3 8B	102.55	64.20	71.82	85.22	114.79	121.29
Llama 3.1 8B	101.43	62.98	70.68	83.83	113.36	120.33
Gemma 7B	460.86	399.22	427.14	468.66	463.11	483.29
Gemma 2 9B	197.41	200.76	192.07	197.88	196.56	202.69
Qwen 2 7B	120.44	83.34	94.87	107.84	133.38	136.52
Qwen 1.5 7B	124.02	89.36	100.55	113.08	135.54	139.13
EMMA-500 Llama 2 7B	68.11	50.12	55.62	64.78	60.53	65.68

Table 10: 3-shot results on FLORES-200 (X-Eng, BLEU/chrF++). EMMA-500 Llama 2 7B has better average performance than all baselines.

Model	Avg	High	Medium-High	Medium	Medium-Low	Low
Llama 2 7B	12.93/ 30.32	19.91/ 39.04	17.56/ 35.84	12.49/ 29.81	8.27/ 24.35	6.96/ 23.36
Llama 2 7B Chat	12.28/ 31.72	18.98/ 39.65	17.06/ 37.03	11.74/ 31.1	7.79/ 26.34	6.18/ 25.03
CodeLlama 2 7B	10.82/ 28.57	17.39/ 37.43	15.27/ 33.94	10.39/ 28.05	6.45/ 22.85	5.04/ 21.48
LLaMAX Llama 2 7B	1.99/ 13.66	3.68/ 22.18	2.95/ 18.15	1.83/ 12.84	0.67/ 7.2	1.01/ 9.04
LLaMAX Llama 2 7B Alpaca	22.29/ 42.27	32.83/ 54.56	30.04/ 51.25	21.7/ 41.94	13.06/ 31.32	14.24/ 32.88
MaLA-500 Llama 2 10B v1‡	2.29/ 13.6	4.64/ 15.95	3.18/ 14.64	2.68/ 14.23	1.24/ 12.58	0.33/ 11.18
MaLA-500 Llama 2 10B v2‡	2.87/ 15.44	5.58/ 18.65	3.81/ 16.33	3.55/ 16.29	1.63/ 14.2	0.55/ 12.76
Yayi Llama 2 7B	12.98/ 31.38	19.48/ 39.58	17.55/ 36.71	12.47/ 30.79	8.54/ 25.63	7.22/ 24.84
TowerBase Llama 2 7B	13.74/ 31.47	21.76/ 40.96	18.92/ 37.27	13.15/ 30.9	8.3/ 25.05	7.21/ 24.1
TowerInstruct Llama 2 7B	4.81/ 25.43	9.18/ 34.4	6.66/ 30.01	4.62/ 25.22	2.64/ 20.24	1.8/ 18.69
Occiglot Mistral 7B v0.1	13.12/ 31.13	19.53/ 38.93	17.57/ 36.27	13.07/ 31.2	9.03/ 26.15	6.86/ 23.83
Occiglot Mistral 7B v0.1 Instruct	11.61/ 31.65	16.72/ 39.28	15.06/ 36.48	11.7/ 31.73	8.48/ 26.88	6.54/ 24.7
BLOOM 7B	9.57/ 27.84	15.75/ 36.65	9.65/ 28.19	9.42/ 27.81	6.81/ 23.95	8.61/ 25.89
BLOOMZ 7B†	20.22/ 34.74	32.23/ 47.03	19.2/ 34.08	20.09/ 34.49	16.25/ 30.58	18.54/ 32.63
mGPT	5.29/ 20.69	9.37/ 26.64	8.28/ 25.29	3.41/ 17.87	2.43/ 16.07	2.84/ 17.28
mGPT-13B	7.42/ 24.58	12.61/ 31.95	11.11/ 30.16	5.72/ 22.49	3.57/ 18.16	4.11/ 20.04
Yayi 7B	4.82/ 21.36	5.69/ 25.18	4.53/ 19.97	4.41/ 21.52	3.71/ 19.18	6.13/ 23.12
Aya 23 8B	13.87/ 32.36	19.87/ 40.18	18.5/ 37.91	12.99/ 31.99	8.19/ 25.6	9.84/ 26.56
Aya expanse 8B	13.12/ 36.86	17.83/ 44.92	15.44/ 41.09	13.14/ 37.16	9.26/ 30.39	10.72/ 31.97
Llama 3 8B	23.78/ 43.72	33.71/ 55.36	30.31/ 51.3	24.75/ 44.91	15.18/ 33.65	16.01/ 34.65
Llama 3.1 8B	24.19/ 44.1	34.15/ 55.7	30.79/ 51.7	24.98/ 45.26	15.89/ 34.24	16.13/ 34.85
Gemma 2 9B	23.15/ 38.87	33.11/ 51.36	30.81/ 48.53	25.58/ 41.23	15.37/ 30.03	11.73/ 24.15
Gemma 7B	23.79/ 43.68	34.23/ 55.77	29.87/ 50.95	24.0/ 44.25	16.16/ 34.36	16.03/ 34.58
Qwen 1.5 7B	15.58/ 35.87	24.07/ 46.29	19.92/ 40.74	15.76/ 36.27	9.74/ 28.81	9.77/ 29.13
Qwen 2 7B	17.39/ 37.61	27.63/ 50.06	22.48/ 43.28	18.13/ 38.63	9.89/ 28.54	10.64/ 29.99
EMMA-500 Llama 2 7B	25.37/ 45.78	32.24/ 53.74	31.39/ 52.85	25.72/ 46.16	20.32/ 39.96	17.18/ 36.15

Table 11: 3-shot results on FLORES-200 (Eng-X, BLEU/chrF++). EMMA-500 Llama 2 7B has better average performance than all baselines.

Model	Avg	High	Medium-High	Medium	Medium-Low	Low
Llama 2 7B	4.62/ 15.13	10.77/ 24.38	8.56/ 21.4	2.55/ 13.72	0.74/ 8.72	0.7/ 7.92
Llama 2 7B Chat	4.95/ 16.95	10.87/ 24.51	8.54/ 22.69	3.25/ 15.5	1.52/ 12.08	0.94/ 10.03
CodeLlama 2 7B	4.27/ 14.94	10.04/ 23.48	7.79/ 20.79	2.57/ 14.2	0.71/ 9.27	0.58/ 7.49
LLaMAX Llama 2 7B	0.8/ 7.42	1.85/ 12.06	1.2/ 9.74	0.54/ 6.55	0.22/ 4.52	0.38/ 4.81
LLaMAX Llama 2 7B Alpaca	12.51/ 28.35	24.8/ 41.76	18.69/ 38.42	10.1/ 27.27	3.79/ 16.53	6.68/ 18.15
MaLA-500 Llama 2 10B v1‡	0.6/ 6.08	1.51/ 9.0	1.13/ 8.19	0.35/ 5.99	0.07/ 4.5	0.02/ 2.9
MaLA-500 Llama 2 10B v2‡	0.54/ 6.38	1.4/ 9.19	1.02/ 8.42	0.24/ 5.99	0.07/ 5.14	0.02/ 3.27
Yayi Llama 2 7B	4.41/ 14.87	10.49/ 24.0	8.21/ 21.27	2.52/ 13.57	0.6/ 8.49	0.53/ 7.42
TowerBase Llama 2 7B	4.83/ 16.03	11.89/ 24.15	8.33/ 21.46	2.57/ 14.49	1.38/ 11.6	0.74/ 8.9
TowerInstruct Llama 2 7B	3.23/ 15.64	7.22/ 22.65	4.99/ 20.0	2.2/ 14.9	1.62/ 12.31	0.73/ 8.97
Occiglot Mistral 7B v0.1	4.32/ 16.1	10.5/ 23.74	6.95/ 20.91	2.87/ 15.44	1.47/ 12.0	0.79/ 9.09
Occiglot Mistral 7B v0.1 Instruct	3.99/ 15.8	9.46/ 23.17	6.46/ 20.73	2.68/ 15.29	1.31/ 11.31	0.84/ 9.04
BLOOM 7B	2.81/ 11.8	7.53/ 19.0	3.12/ 13.36	2.05/ 11.48	0.85/ 8.0	2.09/ 9.22
BLOOMZ 7B†	7.44/ 16.1	23.64/ 32.22	7.46/ 16.62	6.98/ 16.05	1.28/ 9.99	4.17/ 11.77
mGPT	2.59/ 12.56	5.24/ 17.04	4.75/ 16.92	1.14/ 9.75	0.78/ 9.24	0.84/ 9.08
mGPT-13B	3.88/ 14.57	8.32/ 21.7	6.84/ 20.55	2.06/ 12.23	0.9/ 8.4	1.33/ 9.58
YaYi 7B	4.37/ 13.5	13.72/ 26.28	4.68/ 14.31	3.35/ 12.89	0.91/ 8.51	2.55/ 10.08
Aya 23 8B	6.46/ 16.15	11.99/ 22.87	9.64/ 21.38	3.57/ 13.86	2.09/ 10.15	5.17/ 12.25
Aya expanse 8B	6.88/ 23.89	12.41/ 31.11	8.92/ 28.5	5.09/ 22.45	3.55/ 18.29	5.32/ 19.53
Llama 3 8B	9.93/ 24.08	20.38/ 36.87	14.95/ 32.05	8.89/ 24.28	2.83/ 14.26	4.2/ 14.29
Llama 3.1 8B	10.11/ 24.69	20.82/ 37.39	15.3/ 32.82	8.85/ 24.85	2.9/ 14.83	4.23/ 14.81
Gemma 2 9B	12.09/ 26.48	24.62/ 40.69	17.82/ 35.51	10.68/ 26.58	3.38/ 15.02	5.94/ 15.98
Gemma 7B	9.05/ 23.05	17.58/ 34.5	13.62/ 30.16	7.96/ 22.85	2.64/ 14.11	4.47/ 14.82
Qwen 1.5 7B	5.87/ 17.77	14.05/ 28.6	8.88/ 23.57	3.85/ 17.07	1.7/ 10.85	2.35/ 10.21
Qwen 2 7B	5.56/ 17.17	13.22/ 27.65	8.21/ 22.36	4.15/ 16.93	1.56/ 10.47	2.19/ 10.11
EMMA-500 Llama 2 7B	15.58/ 33.25	26.37/ 42.4	21.96/ 41.98	13.4/ 32.06	9.15/ 27.99	7.92/ 21.25

Table 12: 3-shot results on SIB-200 and Taxi-1500 (ACC %). EMMA-500 Llama 2 7B has better average performance than Llama 2 models and comparable performance with multilingual LLMs, and has comparable performance with the compared LLMs.

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	26.64	24.69	22.05	19.68	19.00	17.54	19.50	19.49	18.47	17.46	17.37
Llama 2 7B Chat	25.58	29.72	28.11	25.01	23.03	21.91	15.44	18.73	17.66	16.61	15.59	15.22
CodeLlama 2 7B	23.35	26.06	25.42	23.10	21.42	20.37	17.03	17.45	17.41	17.20	17.05	17.00
LLaMAX Llama 2 7B	10.61	12.42	11.60	10.01	9.45	9.54	23.52	23.20	23.40	23.76	23.56	23.52
LLaMAX Llama 2 7B Alpaca	27.89	33.09	32.12	27.16	23.38	22.82	15.09	18.70	16.88	15.99	15.00	14.91
MaLA-500 Llama 2 10B v1	23.25	23.30	23.64	22.88	22.76	23.58	25.27	23.90	24.02	24.76	24.57	25.43
MaLA-500 Llama 2 10B v2	19.30	18.93	21.05	19.49	17.55	18.46	23.39	21.36	22.30	21.32	21.72	23.67
Yayi Llama 2 7B	24.57	29.04	27.17	24.42	21.44	20.69	17.73	18.74	18.46	18.19	17.89	17.65
TowerBase Llama 2 7B	19.34	22.00	20.92	18.74	17.90	16.93	17.73	18.49	18.81	18.67	18.10	17.61
TowerInstruct Llama 2 7B	20.53	23.21	21.96	20.26	19.15	18.04	17.29	20.17	19.60	18.08	17.40	17.09
Occiglot Mistral 7B v0.1	32.69	38.80	35.82	31.74	28.92	28.36	22.26	24.64	22.91	22.99	22.33	22.15
Occiglot Mistral 7B v0.1 Instruct	34.31	39.48	37.16	33.36	31.47	30.08	18.76	24.30	20.90	19.41	19.18	18.48
BLOOM 7B	17.81	23.13	18.05	17.17	15.76	17.02	14.76	15.58	14.89	14.56	15.11	14.73
BLOOMZ 7B	29.73	30.39	29.63	29.80	29.53	29.70	16.96	16.93	16.98	16.96	16.99	16.95
mGPT	27.11	28.58	27.99	26.73	25.89	26.48	10.72	8.67	8.44	9.92	10.29	10.93
mGPT-13B	33.20	36.69	34.27	34.48	29.39	32.26	17.23	17.98	16.44	15.88	16.10	17.38
Yayi 7B	35.76	40.57	36.20	35.63	34.72	33.18	16.12	16.65	16.38	16.45	15.83	16.11
Aya 23 8B	41.50	46.78	45.27	42.43	35.27	37.87	22.64	22.95	22.50	22.13	22.41	22.67
Aya expanse 8B	57.01	65.05	61.51	56.84	48.59	53.80	18.73	20.02	18.60	18.93	19.39	18.66
Llama 3 8B	63.69	73.45	70.25	64.62	53.16	56.96	21.73	31.84	27.08	25.60	22.61	21.05
Llama 3.1 8B	61.42	70.70	67.90	61.99	51.46	54.75	20.20	27.51	25.21	24.43	20.97	19.59
Gemma 7B	58.21	68.06	64.55	58.16	48.86	51.12	18.05	28.68	28.55	24.99	20.28	16.89
Gemma 2 9B	46.25	51.77	49.00	46.92	43.04	40.45	13.83	24.29	24.13	18.74	14.97	12.83
Qwen 1.5 7B	47.95	56.00	51.81	48.25	41.56	42.86	7.29	12.65	11.45	8.78	7.30	6.92
Qwen 2 7B	54.95	66.37	60.14	55.17	45.19	49.25	21.87	27.37	25.57	24.01	22.33	21.47
EMMA-500 Llama 2 7B	31.27	32.75	33.28	30.99	30.83	27.60	19.82	23.66	23.33	22.77	22.00	19.30

Table 13: 0-shot results (ACC %) on commonsense reasoning: XCOPA, XStoryCloze, and XWinograd. EMMA-500 Llama 2 7B has better average performance than Llama 2 models and multilingual LLMs on XCOPA, XStoryCloze, and comparable performance on XWinograd.

Model	XCOPA				XStoryCloze				XWinograd
	Avg	High	Medium	Low	Avg	High	Medium	Low	Avg
Llama 2 7B	56.67	62.10	53.90	51.60	57.55	63.38	54.45	49.21	72.47
Llama 2 7B Chat	55.85	61.25	53.13	50.60	58.41	64.80	54.77	49.74	69.45
CodeLlama 2 7B	54.69	58.70	52.53	51.60	55.68	60.68	52.33	49.90	70.92
LLaMAX Llama 2 7B	54.38	55.50	54.13	51.40	60.36	64.34	58.82	53.47	67.49
LLaMAX Llama 2 7B Alpaca	56.60	59.80	55.17	52.40	63.83	69.08	62.01	54.33	69.86
MaLA-500 Llama 2 10B v1	53.09	53.55	53.27	50.20	53.07	58.15	49.22	48.08	65.89
MaLA-500 Llama 2 10B v2	53.09	53.55	53.27	50.20	53.07	58.15	49.22	48.08	65.89
YaYi Llama 2 7B	56.71	62.10	54.13	50.60	58.42	64.98	54.91	49.04	74.50
TowerBase Llama 2 7B	56.33	62.50	52.90	52.20	57.78	64.35	53.67	49.57	74.29
TowerInstruct Llama 2 7B	57.05	62.90	54.00	52.00	59.24	66.83	54.53	49.70	74.00
Occiglot Mistral 7B v0.1	56.67	62.80	53.37	52.00	58.10	65.18	53.28	50.03	74.61
Occiglot Mistral 7B v0.1 Instruct	56.55	62.85	52.97	52.80	59.39	66.94	54.33	50.63	72.93
BLOOM 7B	56.89	59.95	55.87	50.80	59.30	61.99	59.05	53.08	70.13
BLOOMZ 7B	54.87	56.35	54.60	50.60	57.12	61.14	55.82	49.67	67.95
mGPT	55.04	57.10	54.40	50.60	54.43	54.96	54.53	52.91	59.69
mGPT 13B	56.18	59.75	55.13	48.20	56.44	57.76	56.35	53.31	63.59
YaYi 7B	56.64	59.55	55.50	51.80	60.67	64.90	59.40	52.65	69.79
Aya 23 8B	55.13	59.05	53.30	50.40	60.93	67.27	58.82	49.34	65.84
Aya expanse 8B	56.38	61.50	53.77	51.60	64.80	73.10	61.93	49.80	71.94
Llama 3 8B	61.71	68.35	59.03	51.20	63.41	68.50	62.03	53.44	76.84
Llama 3.1 8B	61.71	69.30	58.80	48.80	63.58	68.66	62.09	53.87	75.52
Gemma 7B	63.64	70.35	61.43	50.00	65.01	69.46	64.49	54.93	77.41
Gemma 2 9B	66.33	73.40	64.27	50.40	67.67	72.47	66.69	57.64	80.07
Qwen 2 7B	60.31	68.65	56.40	50.40	61.46	69.45	56.97	50.50	76.44
Qwen 1.5 7B	59.44	66.85	55.90	51.00	59.85	66.62	56.04	50.56	72.59
EMMA-500 Llama 2 7B	63.11	66.60	62.57	52.40	66.38	68.92	65.73	61.32	72.80

Table 14: 0-shot results on XNLI (ACC %).

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
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
mGPT	40.51	42.97	39.65	37.30
mGPT 13B	41.59	44.14	40.42	38.82
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
Llama 3 8B	44.97	48.82	43.84	39.56
Llama 3.1 8B	45.62	49.61	44.04	40.83
Gemma 7B	42.58	46.44	41.00	38.01
Gemma 2 9B	46.74	48.50	45.11	46.49
Qwen 2 7B	42.77	47.31	41.35	36.53
Qwen 1.5 7B	39.47	40.95	38.80	37.83
EMMA-500 Llama 2 7B	45.14	46.09	44.40	44.71

Table 15: Results on XL-Sum (ROUGE-L/BERTScore %).

Model	Avg	High	Medium-High	Medium	Medium-Low	Low
Llama 2 7B	7.11/66.52	7.35/66.79	6.65/66.28	6.50/64.67	9.73/61.90	8.22/71.32
Llama 2 7B Chat	8.84/68.44	9.48/68.48	8.27/68.15	7.57/66.82	12.14/64.16	10.18/73.54
CodeLlama 2 7B	7.15/65.74	7.28/65.08	6.80/65.90	6.37/62.96	9.93/63.54	8.26/71.32
LLaMAX Llama 2 7B	5.29/64.59	5.07/64.83	5.19/64.71	4.94/61.43	7.19/61.85	5.94/69.14
LLaMAX Llama 2 7B Alpaca	10.11/69.24	10.42/69.70	9.41/68.68	9.45/68.14	12.65/64.86	12.15/73.80
MaLA-500 Llama 2 10B v1	5.45/63.96	5.97/64.03	5.21/63.88	4.77/61.12	7.05/63.83	5.59/68.16
MaLA-500 Llama 2 10B v2	5.44/64.28	5.86/64.45	5.22/64.07	4.86/61.35	6.86/64.33	5.60/68.82
Yayi Llama 2 7B	7.98/67.21	8.67/67.39	7.48/67.29	7.05/65.59	10.95/60.55	8.55/71.44
TowerBase Llama 2 7B	7.65/67.09	7.92/67.26	7.25/67.06	6.64/65.02	9.90/63.33	9.16/71.21
TowerInstruct Llama 2 7B	8.89/68.46	9.44/68.60	8.59/68.51	7.64/66.28	11.90/63.71	9.45/72.89
Occiglot Mistral 7B v0.1	7.33/66.20	7.53/66.40	6.82/65.87	6.80/64.31	9.71/62.74	8.77/71.11
Occiglot Mistral 7B v0.1 Instruct	8.31/66.96	9.03/66.81	7.77/67.07	7.93/63.86	10.41/63.58	8.59/72.52
BLOOM 7B	6.99/64.78	6.78/65.06	6.97/64.77	6.12/61.89	9.79/58.16	7.61/70.93
BLOOMZ 7B	11.15/69.82	13.92/68.73	8.92/69.99	10.24/69.05	12.03/66.18	14.86/74.12
mGPT	3.85/59.74	4.82/59.91	3.60/60.10	2.35/56.02	6.81/59.22	3.67/63.30
mGPT-13B	4.96/62.20	5.85/62.02	4.59/62.54	3.80/58.97	7.92/59.13	4.90/66.99
Yayi 7B	12.06/69.74	13.98/68.93	10.50/69.49	11.05/68.91	12.45/65.87	15.29/75.26
Aya 23 8B	8.68/66.79	10.03/67.66	8.24/66.26	6.80/68.73	12.63/63.26	8.47/65.68
Aya expanse 8B	10.51/68.73	11.61/68.56	9.66/68.54	10.21/71.02	12.88/65.38	10.90/68.02
Llama 3 8B	8.47/67.08	8.06/66.79	8.10/66.95	8.36/64.85	9.65/65.08	10.52/72.20
Llama 3.1 8B	8.57/66.97	8.01/66.41	8.40/66.84	8.48/65.21	9.05/65.80	10.45/71.68
Gemma 2 9B	7.38/65.45	7.24/64.00	6.87/65.81	7.47/62.95	9.22/63.57	8.89/71.46
Gemma 7B	6.70/64.52	6.86/63.48	6.24/64.91	6.63/67.21	6.73/59.86	8.30/63.35
Qwen 1.5 7B	9.58/69.13	10.61/68.78	8.58/68.89	9.61/67.83	10.79/64.92	10.74/74.40
Qwen 2 7B	10.18/69.35	11.58/69.36	9.31/68.98	9.63/68.03	10.64/64.82	11.15/74.47
EMMA-500 Llama 2 7B	8.58/67.20	8.22/65.87	8.24/67.26	8.00/69.48	11.47/65.55	10.37/67.41

Table 16: 3-shot results (ACC %) on MGSM obtained by direct and CoT prompting.

Model	Direct Prompting				CoT Prompting			
	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
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
mGPT	1.35	1.67	0.53	0.00	1.42	1.93	0.67	0.00
mGPT 13B	1.31	1.80	0.67	0.00	1.53	1.67	1.20	0.00
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
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
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 2 7B	48.95	54.40	38.80	14.80	51.47	58.93	39.07	14.40
Qwen 1.5 7B	31.56	40.00	16.00	4.00	30.36	40.60	14.80	2.40
EMMA-500 Llama 2 7B	17.02	19.20	11.87	2.40	18.09	20.00	13.20	2.80

Table 17: 0-shot results (ACC %) on BELEBELE.

Model	Avg	High	Medium-High	Medium	Medium-Low	Low
Llama 2 7B	26.27	26.76	26.35	26.07	26.41	26.27
Llama 2 7B Chat	29.05	31.84	29.97	28.95	29.47	29.09
CodeLlama 2 7B	27.38	27.37	27.33	27.30	27.30	27.38
LLaMAX Llama 2 7B	23.09	23.23	23.15	23.07	23.10	23.08
LLaMAX Llama 2 7B Alpaca	24.48	25.46	24.82	24.41	24.60	24.49
MaLA-500 Llama 2 10B v1	22.96	23.02	22.98	22.97	22.98	22.97
MaLA-500 Llama 2 10B v2	22.96	23.02	22.98	22.97	22.98	22.97
YaYi Llama 2 7B	28.32	29.64	28.67	28.11	28.37	28.26
TowerBase Llama 2 7B	26.36	27.43	26.85	26.29	26.48	26.34
TowerInstruct Llama 2 7B	27.93	29.88	28.51	27.57	28.19	27.92
Occiglot Mistral 7B v0.1	30.16	32.25	30.94	30.02	30.40	30.15
Occiglot Mistral 7B v0.1 Instruct	32.05	34.14	32.62	31.74	32.40	32.08
BLOOM 7B	24.11	24.25	24.52	24.12	24.11	24.08
BLOOMZ 7B	39.32	45.43	43.67	41.51	40.08	39.51
mGPT	23.96	24.01	23.81	24.00	23.94	23.98
YaYi 7B	37.97	44.37	42.71	40.49	38.72	38.09
Aya 23 8B	40.08	43.85	41.71	39.22	40.93	39.81
Aya expanse 8B	46.98	52.22	48.99	46.57	48.36	46.93
Llama 3 8B	40.73	46.07	42.92	41.03	41.87	40.88
Llama 3.1 8B	45.19	52.50	48.01	45.65	46.74	45.34
Gemma 7B	43.37	52.63	47.83	44.82	45.43	43.94
Gemma 2 9B	54.49	64.10	58.90	55.83	56.85	55.05
Qwen 2 7B	49.31	57.62	52.20	50.04	51.16	49.48
Qwen 1.5 7B	41.83	48.86	44.79	42.18	43.00	41.78
EMMA-500 Llama 2 7B	26.75	28.32	28.18	27.58	27.14	26.94

Table 18: 5-shot results (ACC %) on ARC multilingual.

Model	Avg	High	Medium	Low
Llama 2 7B	27.5608	33.1244	27.3065	21.0196
Llama 2 7B Chat	28.0155	33.6893	27.7891	21.2910
CodeLlama 2 7B	25.2283	28.8639	24.6368	21.6450
LLaMAX Llama 2 7B	26.0905	30.0022	25.9174	21.4821
LLaMAX Llama 2 7B Alpaca	31.0602	36.8898	31.8518	22.4866
MaLA-500 Llama 2 10B v1	21.1612	21.9174	20.4836	21.3172
MaLA-500 Llama 2 10B v2	21.1612	21.9174	20.4836	21.3172
YaYi Llama 2 7B	28.3975	34.2968	28.3460	21.1068
TowerBase Llama 2 7B	27.9355	35.3215	26.8248	20.5081
TowerInstruct Llama 2 7B	30.1019	38.8803	28.8544	21.1562
Occiglot Mistral 7B v0.1	29.7667	38.3930	28.5078	21.0296
Occiglot Mistral 7B v0.1 Instruct	30.8844	40.2927	29.6522	21.1264
BLOOM 7B	23.6501	26.2685	22.7178	21.8920
BLOOMZ 7B	23.9497	26.9360	22.7439	22.1762
mGPT	20.2422	20.1129	19.6471	21.3709
mGPT 13B	21.7584	22.9872	21.1366	21.2329
YaYi 7B	24.4413	27.9617	23.2904	21.9109
Aya 23 8B	31.0838	40.0507	30.0173	21.6080
Aya expanse 8B	36.5617	47.8696	36.1507	23.0948
Llama 3 8B	34.7957	42.4323	35.5256	24.0638
Llama 3.1 8B	34.9336	42.4323	35.8939	23.9995
Gemma 7B	38.6827	46.4619	40.4660	26.0609
Gemma 2 9B	44.1534	54.5890	46.1756	27.8228
Qwen 2 7B	33.8204	43.8754	32.6423	23.1660
Qwen 1.5 7B	28.9263	35.5527	28.1392	21.9223
EMMA-500 Llama 2 7B	29.5282	34.1004	29.8165	23.3444

Table 19: Results on Multipl-E. For language-level breakdowns, refer to Tables 23 to 25 in the appendix.

Model	Avg Pass@1	Avg Pass@10	Avg Pass@25
Llama 2 7B	8.92%	19.45%	25.68%
CodeLlama 2 7B	28.43%	50.83%	63.92%
LLaMAX 2 7B	0.35%	1.61%	2.67%
MaLA-500 Llama 2 10B V2	0.0%	0.0%	0.0%
TowerBase Llama 2 7B	3.61%	6.65%	8.97%
Occiglot Mistral 7B v0.1	21.26%	31.37%	45.86%
Bloom 7B	5.34%	10.49%	14.65%
BloomZ 2 7B	5.85%	11.40%	15.76%
Aya23 8B	9.19%	23.52%	32.09%
Mistral 7B v0.3	26.10%	48.68%	59.05%
Llama 3 8B	30.09%	53.82%	64.01%
LLaMAX 3 8B	3.00%	7.23%	10.67%
Gemma 7B	28.55%	54.27%	64.75%
CodeGemma 7B	31.51%	63.13%	72.65%
Qwen 1.5 7B	21.05%	37.19%	47.63%
Qwen 2 7B	38.68%	62.63%	71.55%
EMMA-500 Llama 2 7B	11.38%	19.02%	26.16%

Table 20: 0-shot results (ACC %) on XCOPA in all languages

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
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
mGPT	55.04	53.00	2.23	49.80	2.24	58.80	2.20	58.20	2.21	50.60	2.24	56.40	2.22	53.20	2.23	55.20	2.23	56.00	2.22	60.20	2.19	54.00	2.23
mGPT 13B	56.18	50.80	2.24	51.80	2.24	62.60	2.17	60.80	2.19	48.20	2.24	58.00	2.21	54.80	2.23	52.80	2.23	56.80	2.22	63.40	2.16	58.00	2.21
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
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
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 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 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
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

Table 21: 0-shot results (ACC %) on XStoryCloze 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
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
mGPT	54.43	49.31	1.29	59.96	1.26	55.46	1.28	54.60	1.28	52.75	1.28	53.14	1.28	51.22	1.29	56.65	1.28	55.00	1.28	57.25	1.27	53.41	1.28
mGPT 13B	56.44	51.62	1.29	64.20	1.23	58.64	1.27	55.39	1.28	55.06	1.28	58.11	1.27	51.22	1.29	59.43	1.26	54.60	1.28	57.64	1.27	54.93	1.28
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
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
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 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 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
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

Table 22: 0-shot results (ACC %) on XWinograd in all languages

Model	Avg	en-acc	stderr	fr-acc	stderr	jp-acc	stderr	pt-acc	stderr	ru-acc	stderr	zh-acc	stderr
Llama 2 7B	72.47	87.91	0.68	66.27	5.22	70.28	1.48	72.24	2.77	68.25	2.63	69.84	2.05
Llama 2 7B Chat	69.45	85.55	0.73	71.08	5.01	68.40	1.50	64.64	2.95	65.71	2.68	61.31	2.17
CodeLlama 2 7B	70.92	84.52	0.75	67.47	5.17	63.61	1.55	72.24	2.77	66.67	2.66	71.03	2.02
LLaMAX Llama 2 7B	67.49	77.89	0.86	61.45	5.37	71.01	1.47	65.40	2.94	55.56	2.80	73.61	1.97
LLaMAX Llama 2 7B Alpaca	69.86	82.75	0.78	66.27	5.22	72.37	1.44	66.16	2.92	59.37	2.77	72.22	2.00
MaLA-500 Llama 2 10B v1	65.89	83.66	0.77	63.86	5.31	60.17	1.58	69.20	2.85	61.90	2.74	56.55	2.21
MaLA-500 Llama 2 10B v2	65.89	83.66	0.77	63.86	5.31	60.17	1.58	69.20	2.85	61.90	2.74	56.55	2.21
YaYi Llama 2 7B	74.50	88.52	0.66	71.08	5.01	72.26	1.45	74.14	2.70	71.75	2.54	69.25	2.06
TowerBase Llama 2 7B	74.29	87.14	0.69	74.70	4.80	69.45	1.49	75.67	2.65	64.76	2.70	74.01	1.96
TowerInstruct Llama 2 7B	74.00	86.28	0.71	77.11	4.64	68.40	1.50	77.19	2.59	66.35	2.67	68.65	2.07
Occiglot Mistral 7B v0.1	74.61	86.54	0.71	79.52	4.46	66.01	1.53	71.86	2.78	66.35	2.67	77.38	1.87
Occiglot Mistral 7B v0.1 Instruct	72.93	85.89	0.72	74.70	4.80	64.55	1.55	70.72	2.81	64.76	2.70	76.98	1.88
BLOOM 7B	70.13	82.24	0.79	71.08	5.01	58.81	1.59	76.81	2.61	57.46	2.79	74.40	1.95
BLOOMZ 7B	67.95	83.40	0.77	73.49	4.87	57.56	1.60	67.30	2.90	54.92	2.81	71.03	2.02
mGPT	59.69	62.67	1.00	59.04	5.43	53.49	1.61	57.41	3.05	58.10	2.78	67.46	2.09
mGPT 13B	63.59	70.62	0.94	63.86	5.31	57.77	1.60	61.22	3.01	60.00	2.76	68.06	2.08
YaYi 7B	69.79	84.04	0.76	75.90	4.72	58.08	1.59	72.24	2.77	55.87	2.80	72.62	1.99
Aya 23 8B	65.84	75.31	0.89	62.65	5.34	65.07	1.54	63.50	2.97	60.63	2.76	67.86	2.08
Aya expanse 8B	71.94	78.58	0.85	75.90	4.72	71.43	1.46	70.34	2.82	64.76	2.70	70.63	2.03
Llama 3 8B	76.84	86.80	0.70	71.08	5.01	75.29	1.39	79.85	2.48	71.43	2.55	76.59	1.89
Llama 3.1 8B	75.52	87.57	0.68	66.27	5.22	75.29	1.39	80.61	2.44	67.62	2.64	75.79	1.91
Gemma 7B	77.41	88.00	0.67	73.49	4.87	75.39	1.39	78.33	2.55	72.06	2.53	77.18	1.87
Gemma 2 9B	80.07	89.42	0.64	75.90	4.72	79.77	1.30	82.89	2.33	74.29	2.47	78.17	1.84
Qwen 2 7B	76.44	86.32	0.71	69.88	5.07	68.82	1.50	78.71	2.53	71.75	2.54	83.13	1.67
Qwen 1.5 7B	72.59	83.31	0.77	69.88	5.07	65.90	1.53	70.34	2.82	66.35	2.67	79.76	1.79
EMMA-500 Llama 2 7B	72.80	82.45	0.79	72.29	4.94	71.64	1.46	69.20	2.85	69.21	2.61	72.02	2.00

Table 23: Pass@1 on Multipl-E.

Model	C++	C#	Java	JavaScript	Python	Rust	TypeScript
Llama 2 7B	6.74%	5.65%	8.54%	11.34%	11.98%	6.12%	12.04%
CodeLlama 2 7B	26.70%	20.44%	30.56%	32.89%	28.76%	26.23%	33.45%
LLaMAX 2 7B	0.0%	0.0%	0.02%	0.68%	0.0%	0.0%	1.78%
MaLA-500 Llama 2 10B V2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TowerBase Llama 2 7B	1.01%	6.85%	7.86%	0.87%	8.68%	0.0%	0.0%
Occiglot Mistral 7B v0.1	23.16%	16.14%	19.83%	26.92%	21.45%	16.34%	24.97%
Bloom 7B	5.25%	2.97%	6.37%	6.76%	7.65%	1.01%	7.34%
BloomZ 2 7B	6.87%	3.59%	6.02%	6.96%	7.33%	2.13%	8.04%
Aya23 8B	16.03%	7.91%	14.29%	6.23%	3.56%	11.43%	4.90%
Mistral 7B v0.3	26.12%	22.87%	25.54%	35.24%	24.91%	19.24%	28.76%
Llama 3 8B	34.12%	21.06%	26.56%	36.12%	30.22%	25.19%	37.34%
LLaMAX 3 8B	0.0%	0.0%	0.0%	9.73%	0.91%	0.21%	10.12%
Gemma 7B	29.66%	21.40%	27.35%	35.29%	30.11%	25.48%	30.57%
CodeGemma 7B	33.91%	21.05%	29.43%	37.78%	32.56%	28.70%	37.14%
Qwen 1.5 7B	22.04%	12.42%	17.68%	27.58%	32.11%	8.32%	27.21%
Qwen 2 7B	43.51%	21.47%	38.95%	46.31%	37.49%	34.61%	48.41%
EMMA-500 Llama 2 7B	11.34%	8.94%	11.93%	11.67%	18.97%	6.86%	9.94%

Table 24: Pass@10 on Multipl-E.

Model	C++	C#	Java	JavaScript	Python	Rust	TypeScript
Llama 2 7B	17.92%	14.22%	20.78%	23.12%	24.86%	13.08%	22.19%
CodeLlama 2 7B	51.39%	37.13%	50.38%	59.08%	56.58%	47.68%	53.60%
LLaMAX 2 7B	0.96%	0.0%	0.53%	3.99%	0.0%	0.12%	5.64%
MaLA-500 Llama 2 10B V2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TowerBase Llama 2 7B	1.01%	6.85%	7.86%	0.87%	8.68%	0.0%	0.0%
Occiglot Mistral 7B v0.1	36.08%	27.49%	35.87%	47.45%	41.76%	29.80%	43.13%
Bloom 7B	11.35%	6.86%	12.79%	12.44%	14.24%	3.68%	12.08%
BloomZ 2 7B	12.55%	8.21%	12.94%	14.10%	14.84%	3.91%	13.27%
Aya23 8B	28.49%	17.34%	27.12%	23.19%	26.72%	26.14%	15.67%
Mistral 7B v0.3	50.23%	35.18%	46.83%	57.23%	54.29%	42.77%	54.22%
Llama 3 8B	55.13%	36.67%	54.34%	62.65%	59.24%	45.68%	63.02%
LLaMAX 3 8B	0.72%	0.0%	0.97%	22.91%	1.58%	1.38%	23.04%
Gemma 7B	55.21%	39.09%	52.02%	61.88%	60.09%	50.34%	61.23%
CodeGemma 7B	62.29%	45.11%	59.74%	70.96%	69.76%	61.27%	72.76%
Qwen 1.5 7B	40.11%	28.95%	37.56%	48.35%	40.34%	20.19%	44.85%
Qwen 2 7B	64.58%	43.17%	63.28%	73.21%	54.34%	65.48%	74.32%
EMMA-500 Llama 2 7B	22.84%	14.29%	21.28%	18.22%	24.94%	15.68%	15.91%

Table 25: Pass@25 on Multipl-E.

Model	C++	C#	Java	JavaScript	Python	Rust	TypeScript
Llama 2 7B	24.89%	18.77%	26.45%	30.31%	31.09%	18.70%	29.55%
CodeLlama 2 7B	64.12%	45.96%	61.98%	72.76%	71.77%	59.98%	70.89%
LLaMAX 2 7B	1.47%	0.0%	1.04%	6.78%	0.0%	0.34%	9.04%
MaLA-500 Llama 2 10B V2	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TowerBase Llama 2 7B	1.86%	15.74%	22.45%	2.53%	20.18%	0.0%	0.0%
Occiglot Mistral 7B v0.1	44.17%	33.89%	42.77%	56.63%	49.87%	37.94%	55.76%
Bloom 7B	17.27%	10.07%	17.23%	17.55%	17.93%	5.70%	16.81%
BloomZ 2 7B	17.41%	11.98%	18.66%	18.37%	18.78%	6.83%	18.31%
Aya23 8B	37.56%	21.87%	36.45%	33.57%	36.09%	35.34%	23.75%
Mistral 7B v0.3	60.93%	44.80%	56.12%	66.34%	65.33%	55.31%	64.51%
Llama 3 8B	64.42%	43.34%	62.14%	73.51%	71.82%	57.78%	75.09%
LLaMAX 3 8B	1.71%	0.0%	2.32%	35.61%	2.77%	2.08%	30.19%
Gemma 7B	66.14%	47.08%	63.24%	70.22%	70.47%	63.52%	72.56%
CodeGemma 7B	73.66%	50.79%	69.96%	79.48%	78.68%	74.87%	81.12%
Qwen 1.5 7B	51.87%	37.69%	48.71%	59.71%	49.69%	29.78%	55.98%
Qwen 2 7B	74.33%	51.87%	72.03%	81.67%	65.38%	74.97%	80.59%
EMMA-500 Llama 2 7B	31.93%	18.69%	29.85%	26.45%	32.55%	21.32%	22.34%

Table 26: 0-shot results (ACC %) on XNLI in all languages.

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
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
mGPT	40.51	33.82	0.95	41.73	0.99	44.78	1.00	35.42	0.96	49.36	1.00	42.81	0.99	44.06	1.00	41.33	0.99	43.21	0.99	41.89	0.99	35.78	0.96	40.12	0.98	34.22	0.95	45.50	1.00	33.61	0.95
mGPT 13B	41.59	33.37	0.95	44.82	1.00	46.47	1.00	37.71	0.97	49.68	1.00	44.50	1.00	45.62	1.00	43.05	0.99	44.86	1.00	40.32	0.98	38.67	0.98	38.84	0.98	37.47	0.97	44.74	1.00	33.69	0.95
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
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
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 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 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
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

Table 27: 3-shot results (ACC %) 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	6.69	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
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
mGPT	1.35	0.00	0.00	2.00	0.89	3.20	1.12	0.80	0.56	2.40	0.97	1.60	0.80	2.40	0.97	1.60	0.80	0.00	0.00	0.00	0.00	0.80	0.56
mGPT 13B	1.31	0.00	0.00	1.60	0.80	1.60	0.80	0.80	0.56	3.60	1.18	1.60	0.80	2.00	0.89	1.20	0.69	0.00	0.00	0.00	0.56	1.20	0.69
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
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
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 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 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
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

Table 28: 3-shot results (ACC %) on MGSM in all languages by CoT 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
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
mGPT	1.42	0.00	0.00	2.00	0.89	2.00	0.89	1.60	0.80	2.40	0.97	1.20	0.69	2.00	0.89	2.80	1.05	0.00	0.00	0.00	0.00	1.60	0.80
mGPT 13B	1.53	0.00	0.00	2.00	0.89	3.60	1.18	1.60	0.80	2.40	0.97	1.20	0.69	1.60	0.80	0.80	0.56	0.40	0.40	0.80	0.56	2.00	0.89
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
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
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 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 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
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