



MEXA: MULTILINGUAL EVALUATION OF ENGLISH-CENTRIC LLMs VIA CROSS-LINGUAL ALIGNMENT

Anonymous authors

Paper under double-blind review

ABSTRACT

English-centric large language models (LLMs) often show strong multilingual capabilities. However, the multilingual performance of these models remains unclear and is not thoroughly evaluated for many languages. Most benchmarks for multilinguality focus on classic NLP tasks, or cover a minimal number of languages. We introduce MEXA, a method for assessing the multilingual capabilities of pre-trained English-centric LLMs using parallel sentences, which are available for more languages than existing downstream tasks. MEXA leverages the fact that English-centric LLMs use English as a kind of pivot language in their intermediate layers. It computes the alignment between English and non-English languages using parallel sentences to evaluate the transfer of language understanding from English to other languages. This alignment can be used to estimate task performance in other languages. We conduct studies using various parallel datasets (FLORES-200 and Bible), models (Llama family, Gemma family, Mistral, and OLMo), and established downstream tasks (Belebele, m-MMLU, and m-ARC). We explore different methods to compute embeddings in decoder-only models. Our results show that MEXA, in its default settings, achieves a statistically significant average Pearson correlation of 0.90 with three established downstream tasks across nine models and two parallel datasets. This suggests that MEXA is a reliable method for estimating the multilingual capabilities of English-centric LLMs, providing a clearer understanding of their multilingual potential and the inner workings of LLMs.

 **Leaderboard** [anonymized url]  **Code** [anonymized url]

1 INTRODUCTION

Most state-of-the-art autoregressive large language models (LLMs) are English-centric, closed-source models such as Claude 3 Opus, GPT-4, and Gemini Pro (Anthropic, 2023; OpenAI et al., 2023; Gemini Team et al., 2023); open-weight models such as Llama 3.1, Gemma 2, and Mixtral (Dubey et al., 2024; Gemma Team et al., 2024b; Jiang et al., 2024); and open-source models such as OLMo (Groeneveld et al., 2024). Except for open-source models, where the data is available and thus the language distribution is transparent, there is confusion regarding the capabilities/language distribution of these LLMs in other languages.

Primarily, the focus in evaluating LLMs has been on developing benchmarks to assess their performance in English. Most benchmarks in multilingual setups consist of classical monolingual NLP tasks such as sequence labeling (Ahuja et al., 2023; Lai et al., 2023a), the automatic translation of popular English benchmarks such as MMLU (Hendrycks et al., 2021) into a limited number of languages (Lai et al., 2023b; OpenAI, 2024), or language-specific benchmarks for languages such as Persian (Ghahroodi et al., 2024), Arabic (Koto et al., 2024), Korean (Son et al., 2024), Turkish (Yüksel et al., 2024), and Chinese (Li et al., 2024c).

Most LLMs are English-centric, either by choice or due to the availability of abundant data sources in English. Either way, for these models to be effective in other languages, it is important that the other languages align with the main language, i.e., English. Given such alignment, English could act as a “*rising tide that raises all ships*,” meaning that improvements in English performance could benefit other languages, especially in tasks such as reasoning (Zhu et al., 2024). Contrarily, if a language does not align well with English, an English-centric LLM may not provide *meaningful coverage* for

054 that language. Indeed, Wendler et al. (2024) have found that for Llama 2 (Touvron et al., 2023b), an
055 English-centric LLM, English could be seen as a kind of “pivot” language, enabling to solve complex
056 semantic tasks in a foreign language through a detour into English. More precisely, they show that
057 Llama 2 was able to decode semantically correct next tokens in the middle layers, assigning higher
058 probabilities to the English tokens than to the foreign version, which is only selected in the upper
059 layers. Zhao et al. (2024) present a hypothesis regarding the middle layers of English-centric LLMs,
060 suggesting that these models use English as a means of reasoning while incorporating multilingual
061 knowledge. Based on their analysis, the number of language-specific neurons in the middle layers
062 decreases within the self-attention mechanism but remains consistent across the layers of the feed-
063 forward structure when processing multilingual queries.

064 In this paper, we introduce MEXA, a method that uses the observation that English-centric LLMs
065 semantically use English as a kind of pivot language in their middle layers to evaluate the actual
066 multilingual coverage of LLMs. This is done by measuring how well the embeddings of parallel
067 sentences in the middle layers of LLMs for non-English languages are aligned with the embedding
068 of their corresponding English. We extensively verify the MEXA estimation of language coverage
069 for each LLM, using Pearson correlation between estimated and actual scores for various tasks. We
070 use two parallel datasets: FLORES-200 (NLLB Team et al., 2022) and Bible (Mayer & Cysouw,
071 2014); nine LLMs: Llama family, Gemma family, Mistral, and OLMo; and three tasks: Bele-
072 bele (Bandarkar et al., 2024), m-MMLU, and m-ARC (Lai et al., 2023b). Our results show that
073 MEXA achieves a promising average Pearson correlation of 0.90 with established downstream tasks
074 across nine models and two parallel datasets. In our study on the calculation of MEXA scores, we
075 conduct multiple design analyses to examine the impact of token-level pooling for the embeddings
076 (i.e., using the last token versus a weighted average) and layer-level pooling in computing alignment
077 scores. While MEXA demonstrates a high correlation across most setups, we find that a weighted
078 average based on tokens, combined with mean pooling, yields the best results.

079 2 BACKGROUND AND RELATED WORK

080

081 We discuss the distribution of pre-training data in LLMs and multilingual evaluation benchmarks
082 in Appendices A.1 and A.2 while focusing on cross-lingual alignment here. Research in the cross-
083 lingual alignment field either aims to uncover the underlying mechanisms of alignment and assess its
084 impact on models and downstream tasks, or attempts to enhance model performance by enforcing
085 alignment before, during, or after the pre-training phase. Most of these papers have focused on
086 encoder-only models, such as XLM-R (Conneau et al., 2020a) and mBERT (Devlin et al., 2019),
087 among others (Hämmerl et al., 2024). In this work, we focus primarily on decoder-only models.

088 **Understanding Alignment.** Ye et al. (2023) show that English-centric models such as Llama 1
089 (Touvron et al., 2023a) not only possess multilingual transfer abilities (after fine-tuning on one
090 source language, they can be applied to other languages) but may even surpass the multilingual
091 transfer abilities of multilingual pre-trained models such as BLOOM (BigScience Workshop et al.,
092 2023). Schäfer et al. (2024) find that GPT-2-style decoder-only models show strong cross-lingual
093 generalization when trained on an imbalanced mix of languages. However, when trained on a bal-
094 anced language set, they do not observe increased performance compared to monolingual settings.
095 Wendler et al. (2024) perform single-token analysis to demonstrate that English-centered LLMs,
096 such as Llama 2, use English semantically as an internal latent language in the middle layers when
097 handling multilingual queries. Zhong et al. (2024) extend this analysis to multiple tokens, also
098 showing that an LLM dominated by both English and Japanese uses both languages as internal la-
099 tent languages. Zhao et al. (2024) explore how LLMs handle multilingualism. They hypothesize
100 that LLMs initially interpret the query, converting multilingual inputs into English for task-solving.
101 In the middle layers, the models rely on English with self-attention mechanisms for reasoning, while
102 employing multilingual knowledge through feed-forward structures. In the final layers, LLMs gen-
103 erate responses consistent with the original query language. Li et al. (2024f) and Li et al. (2024b)
104 are even more closely related to ours. Li et al. (2024f) employs absolute cosine similarity values
105 between last token embeddings derived from parallel sentences with English to predict the ranking
106 of language performance across various models. However, as we discuss in Section 3, relying solely
107 on absolute cosine values can be misleading, and as shown in Section 5.3, absolute cosine values
are less correlated with downstream tasks than MEXA score. Li et al. (2024b) uses English probing
tasks and their automatic translations to construct a multilingual evaluation. While they compare

embedding similarity scores between high- and low-resource languages with corresponding evaluation results, similar to Li et al. (2024f), they do not assess whether these correlations hold across other downstream tasks. In Section 5, we demonstrate that MEXA scores align closely with a broad range of downstream tasks.

Boosting Alignment. The idea to enforce alignment in encoder-only models using parallel sentences dates back to (Conneau & Lample, 2019), and has been explored under various guises e.g., using mixed-language sentences and/or bilingual dictionaries Huang et al. (2019); Conneau et al. (2020b); Cao et al. (2020); Kulshreshtha et al. (2020); Efimov et al. (2023); Zhang et al. (2023b). Recently, Li et al. (2024d) improve multilingual alignment by initializing the decoder-only models to generate similar representations of aligned words using contrastive learning and preserves this alignment using a code-switching strategy during pretraining. Liu et al. (2024a) propose a data allocation technique to select a core subset of languages for fine-tuning, better aligning the multilingual capabilities of decoder-only LLMs and making them more truthful in their responses. Li et al. (2024a) propose aligning internal sentence representations across different languages using multilingual contrastive learning and aligning outputs by following cross-lingual instructions in the target language for decoder-only models.

3 MEXA

We now describe the MEXA method for computing the alignment score of language L_1 with respect to the pivot language L_2 , given the language model m . In this paper, we use the term *cross-lingual alignment*, *geometric alignment*, or simply *alignment* to refer to the semantic similarity of multilingual embeddings across languages. L_2 , for English-centric LLMs and in this paper, is English. To assess alignment, we use parallel sentences in two languages, L_1 and L_2 .

What defines semantic similarity in multilingual embeddings across languages? The goal of semantic similarity is to ensure that parallel sentences have sufficiently high similarity, reflecting alignment between the two languages. However, considering only the absolute cosine similarity value as the alignment score does not guarantee proper alignment. For some languages, even non-parallel sentences exhibit similarity scores comparable to those of parallel sentences (see §5.3). This is largely due to the high anisotropy observed in Transformer models, which can lead to so-called hubness issues, making it difficult to distinguish between similar and dissimilar embeddings (Ethayarajh, 2019), especially in multilingual models (Hämmerl et al., 2023; Rajaei & Pilehvar, 2022). However, a direct comparative analysis of the cosine similarity between parallel and non-parallel sentence pairs across languages can help overcome these issues. Instead of using the absolute cosine similarity value for alignment, we assign binary values (1 or 0) based on whether a criterion for semantic similarity is satisfied. This criterion imposes that (a) parallel sentences should have high cosine similarity, and (b) non-parallel pairs should also have low similarity values, ensuring the similarity is not random or biased. Specifically, if the cosine similarity for a pair of parallel sentences is higher than for any non-parallel sentences, we assign a value of 1 for this pair; otherwise, a value of 0. This approach sidesteps the hubness problem since the absolute cosine similarity values themselves are not directly used.

To compute MEXA, we first apply the cosine similarity function to the pairs of embeddings of parallel sentences in languages L_1 and L_2 . In Section 3.1, we describe how embeddings can be computed for each layer l of the autoregressive language model m . We generate a square matrix $C(L_1, L_2, m, l)$ representing cosine similarities of embeddings at the output of layer l for all parallel sentences in languages L_1 and L_2 . We denote $c_{ij}(l)$ the element in the i -th row and j -th column of $C(L_1, L_2, m, l)$. It represents the cosine similarity between the i -th sentence of L_1 and the j -th sentence of L_2 at layer l of language model m . The diagonal elements of C , denoted $c_{ii}(l)$, represent the cosine similarity between parallel sentence pairs from L_1 and L_2 . We define the MEXA alignment score ($\mu(\cdot)$) as follows:

$$\mu(C(L_1, L_2, m, l)) = \frac{1}{n} \sum_{i=1}^n \mathbf{1} \left(c_{ii}(l) > \max_{j \in \{1, \dots, n\} \setminus \{i\}} \{c_{ij}(l), c_{ji}(l)\} \right) \quad (1)$$

where n is the number of diagonal elements (i.e., the dimension of the matrix), and $\mathbf{1}(\cdot)$ is the indicator function, which equals 1 if its argument condition evaluates to true and 0 otherwise. This

alignment score measures how often $c_{ii}(l)$ is the maximum value in both its row and column. The MEXA alignment score can alternatively be understood as a measure of sentence retrieval performance (Hu et al., 2020; Liu et al., 2024b; Hämmerl et al., 2024), with the metric of P@1 applied with queries in language L_1 and answers in L_2 , and vice versa. We discuss other ways to calculate semantic similarity between languages in Appendix A.3.

Layer-wise Pooling. The MEXA alignment score $\mu(C(L_1, L_2, m, l))$ is computed for language L_1 respect to pivot language L_2 for each layer l of the language model m . To compute a single MEXA alignment score given the language model m and L_1, L_2 , we use mean and max pooling strategies over multiple layers.

3.1 SENTENCE EMBEDDINGS

Sentence embeddings are a transformation of a sentence into a fixed-size vector that captures its meaning. The process of computing sentence embeddings can vary depending on the model architecture. Typically, sentence embeddings are used in encoder-based models such as BERT, which employ bidirectional attention. In these models, the hidden states of each token are first extracted, then aggregated, commonly by averaging the hidden states from the output layer. Since attention in these models is bidirectional, each token contributes equally to the final embedding. Alternatively one can use the output of the special [CLS] token as per the original BERT work (Devlin et al., 2019).

In this paper, we focus on autoregressive language models that use a decoder-only architecture. In this architecture, attention is not bidirectional; instead, it takes the form of causal attention (left-to-right). In bidirectional attention, each token has access to every other token in the sequence. However, in causal attention, the embedding of a token at position t is only influenced by the embedding of preceding tokens at positions $0, 1, \dots, t - 1$. Therefore, simple averaging biases the embeddings towards sentence-initial words. Instead, two alternative methods are considered: using only the last token and weighted averaging. We use and compare both methods to acquire the sentence embeddings needed for MEXA.

A standard way to compute a sentence embedding uses only the last token of that sentence. Jiang et al. (2023b) show that using the last token in the format of a prompt template for a sentence s , such as 'This sentence: {s} means in one word:', can be effective. Inspired by this, Li & Li (2024) used the prompt 'Summarize sentence {s} in one word:' to obtain the last token embedding as the sentence-level text embedding. However, not all models are instruction-tuned; some earlier works, such as Neelakantan et al. (2022); Wang et al. (2024); Ma et al. (2024), use the last token without any prompt. Since the models studied in this paper are only pre-trained and use multiple languages in the input, we decided to use the last token method without any preceding instruction.

An alternative is weighted averaging. It relies on the intuition that using only the last token might not represent the entire sentence, as the influence of earlier tokens may have diminished. This suggests that tokens at the end of the sentence should contribute more to the overall embedding than those at the beginning. Another motivation is that sentence-final tokens are influenced by preceding tokens and contain more context, while the representation of sentence-initial tokens has significantly less contextual representation. To address this, Muennighoff (2022) proposes to assign weights to each token based on its position. Thus, the sentence embedding of layer l using position-weighted averaging is:

$$e_l = \sum_{t=1}^T w_t h_{lt} \quad \text{with} \quad w_t = \frac{t}{\sum_{k=1}^T k} \quad (2)$$

where T is the number of tokens in the given sentence, h_{lt} is the embedding of the t -th token at layer l , and e_l is the sentence embedding at layer l .

4 EXPERIMENTS

We conduct experiments using various multi-parallel datasets (FLORES-200 and the Bible), models (Llama family, Gemma family, Mistral, and OLMo), and existing benchmarks/tasks (Belebele, m-

MMLU, m-ARC). Our objective is to assess how well the MEXA alignment score from various parallel datasets correlates with the different benchmarks/tasks for different models.

4.1 PARALLEL DATA

We calculate the MEXA score using the parallel datasets of FLORES-200 (NLLB Team et al., 2022) and the Bible (Mayer & Cysouw, 2014). While there are other high-quality parallel datasets, such as NTREX-128 (Federmann et al., 2022), IN22 (Gala et al.), OPUS-100 (Zhang et al., 2020), Europarl (Koehn, 2005), OpenSubtitles (Lison & Tiedemann, 2016), TED2020 (Reimers & Gurevych, 2020), and Tatoeba (Tatoeba Community, 2006), we specifically chose FLORES-200 due to its high quality and support for a wide range of languages, while the Bible dataset was selected for its extensive language coverage.

FLORES-200 is a parallel corpus, where the English subset is sourced from Wikimedia projects. Each English sentence has been translated into 204 distinct language-script combinations, these translations have been verified by humans. The dataset contains 997 sentences for development, 1012 sentences for dev-test, and 992 sentences for testing. As the FLORES-200 test set is not publicly accessible, we use the dev-test set as our FLORES parallel test corpus, in line with previous studies. For faster computation, we only consider the first 100 sentences from each language. As shown in Appendix A.4, the odds of the MEXA score randomly achieving a high value with 100 sentences are very slim.

The Parallel Bible (Mayer & Cysouw, 2014) covers a very large number of languages. From this resource, we managed to create a subcorpus, a super parallel dataset of the Bible, with 1,401 language-script labels, each containing 103 sentences (i.e., Bible verses).¹ This corpus includes many low-resource languages, many of which are not covered by existing language technologies (Joshi et al., 2020), and MEXA can be adopted since only parallel data is needed. We use all the 103 sentences from each language.

4.2 MODELS

For our experiments, we select models with around 7B parameters, which are considered a base size in the LLM community. The state-of-the-art open-weight models in this range include Llama 1, 2, 3, and 3.1 (Touvron et al., 2023a;b; Meta, 2024; Dubey et al., 2024), Gemma 1 and 2 (Gemma Team et al., 2024a;b), Mistral 0.3 (Jiang et al., 2023a), and the open-source model OLMo 1.7 (Groeneveld et al., 2024). We also select a larger model, Llama 3.1 70B, to show that our findings hold even when scaling further. To apply MEXA, we need to access model weights to compute input sentence embeddings for each layer. We use three popular open-weight model families: Llama, Gemma, and Mistral. As a less multilingual version of state-of-the-art LLMs, we include OLMo, which is trained on a more English-centric corpus of Dolma (Soldaini et al., 2024). Although there are multilingual models such as PolyLM (with support of 18 languages), XGLM (Lin et al., 2022) (with support of 30 languages) and BLOOM (BigScience Workshop et al., 2023) (with support of 46 languages) our focus here is on LLMs which are state-of-the-art in English based tasks such as MMLU (Stanford CRFM, 2024).

4.3 BENCHMARKS

Among the existing evaluation benchmarks in Table 4 from Appendix A.2, we chose the Belebele benchmark (Bandarkar et al., 2024), m-ARC (Lai et al., 2023b), and m-MMLU (Lai et al., 2023b), which support the highest number of high-, medium-, and low-resource languages and are directly related to natural understanding tasks, which is the primary focus of this paper.

Belebele is a multiple-choice reading comprehension task designed to assess language models across a range of high-, medium-, and low-resource languages. Each question in the dataset is accompanied by four possible answers and is linked to a brief passage from the FLORES-200 dataset (NLLB Team et al., 2022). The human annotation process was carefully curated to generate questions that effectively differentiate between various levels of language comprehension, supported by rigorous quality checks. Belebele supports 122 distinct labels (language-script combinations) corresponding

¹[anonymized url]

		Gemma 2 9B	Gemma 1 7B	Llama 3.1 70B	Llama 3.1 8B	Llama 3 8B	Llama 2 7B	Llama 1 7B	Mistral 0.3 7B	OLMo 1.7 7B	AVG
Task _{eng}	Belebele	<u>0.9178</u>	0.8467	0.9456	0.8767	0.8689	0.4822	0.4156	0.8389	0.7711	0.7737
	m-MMLU	<u>0.6998</u>	0.6138	0.7700	0.6315	0.6294	0.4523	0.3569	0.5988	0.5210	0.5859
	m-ARC	<u>0.6775</u>	0.5870	0.7014	0.5794	0.5836	0.5128	0.5000	0.5862	0.4872	0.5795
Task _{L\{eng}}	Belebele (avg., L = 116)	0.7093	0.5633	0.7684	0.5705	0.5533	0.3028	0.2755	0.4457	0.3627	0.5057
	m-MMLU (avg., L = 33)	<u>0.5582</u>	0.4734	0.6384	0.4720	0.4664	0.3260	0.2807	0.4207	0.3390	0.4416
	m-ARC (avg., L = 31)	<u>0.4779</u>	0.4220	0.5054	0.3941	0.3892	0.3174	0.2970	0.3662	0.2731	0.3825
FLORES	μ_{Mean} (avg., L = 116)	0.5088	0.3815	<u>0.4110</u>	0.3963	0.3939	0.0866	0.1946	0.2642	0.0413	0.2976
	μ_{Max} (avg., L = 116)	<u>0.7194</u>	0.5872	0.7725	0.6538	0.6520	0.2464	0.3579	0.4716	0.1965	0.5175
Bible	μ_{Mean} (avg., L = 101)	0.3568	0.2152	<u>0.3169</u>	0.2103	0.2026	0.1246	0.0908	0.1198	0.0121	0.1832
	μ_{Max} (avg., L = 101)	<u>0.6076</u>	0.4021	0.6599	0.4212	0.4190	0.2724	0.2357	0.2606	0.0319	0.3678

Table 1: μ_{pooling} indicates the MEXA score for each corresponding pooling strategy. The embeddings are computed using weighted average based on token positions (Eq. 2). Top values are in **bold**, with second-best underlined.

to 115 distinct languages. However, FLORES-200 does not support 5 of these labels, corresponding to Romanized versions of 5 Indic languages. Therefore, we conducted our analysis between the FLORES-200 and Belebele benchmarks on 117 common labels. Additionally, there are 102 common labels between the Bible parallel data and Belebele benchmark.

Both ARC Challenge (Clark et al., 2018) and MMLU (Hendrycks et al., 2021) are also set up as multiple-choice question-answering tasks, but they focus on different types of knowledge and reasoning skills. ARC Challenge is classified as a common-sense reasoning task, consisting of grade-school level science questions, while MMLU consists of questions across a wide range of subjects, including humanities, social sciences, and more. Lai et al. (2023b) used GPT-3.5-turbo (OpenAI, 2022) and a translation prompt to translate examples from both datasets and create m-ARC and m-MMLU in 31 languages (excluding English). Later, m-MMLU was expanded to also include Icelandic (isl_Latn) and Norwegian (nob_Latn). The Icelandic portion was translated using the Mideind.is, while Norwegian was generated with DeepL.com.² m-MMLU consists of 277 questions in its training set, 13,258 in the test set and 1,433 in the validation set. m-ARC consists of 1,116 questions in the training set, 1,169 in the test set, and 298 in the validation set. We use the entire test set for each of these benchmarks to evaluate the models, except in one case. For Llama 3.1 70B, we use the first 500 questions of m-MMLU instead of the whole set due to resource constraints. Since the selected LLMs used in our experiment are not instruction-tuned, we use 5-shot in-context learning with the lm-evaluation-harness framework, employing log-likelihood-based multiple-choice scoring. Other settings, such as prompt templates, are configured according to the framework’s default (Gao et al., 2023; Biderman et al., 2024).

4.4 EVALUATION MEASURES

We use Pearson correlation to assess the strength of the correlation between MEXA and downstream performance on our evaluation benchmarks. Pearson correlation is a statistical measure that calculates the strength and direction of the linear relationship between two variables. A high Pearson correlation would indicate that MEXA provides a reliable assessment of multilingual capabilities in English-centric LLMs.

5 RESULTS

Table 1 presents the downstream performance of the selected models across three benchmarks, along with MEXA scores from two parallel datasets. Notably, among models with parameter sizes ranging from 7 to 9 billion, both Gemma 2 and Llama 3.1 outperform the others in terms of non-English downstream performance and MEXA scores. The Llama 3.1 and Llama 3 models exhibit similar alignment and downstream task performance; yet, both represent substantial advancements compared to Llama 2. Moreover, results for the Llama 3.1-70B model indicate that scaling can significantly enhance alignment when compared to its smaller version. Interestingly, while Mistral achieves comparable results to Gemma 1 on English benchmarks, it demonstrates inferior alignment, which likely accounts for its reduced performance on non-English tasks. Furthermore, the Llama 2 model achieves higher MEXA scores than OLMo, indicating better alignment. However,

²https://huggingface.co/datasets/alexandrainst/m_mmlu

		Gemma 2 9B	Gemma 1 7B	Llama 3.1 70B	Llama 3.1 8B	Llama 3 8B	Llama 2 7B	Llama 1 7B	Mistral 0.3 7B	OLMo 1.7 7B	AVG	
324 325 326 327 328 329	FLORES weighted average	ρ (μ_{Mean} , Belebele)	0.9247	0.9421	0.8291	0.9478	0.9588	0.8364	0.8404	0.9732	0.8425	0.8994
		ρ (μ_{Max} , Belebele)	0.9623	0.9676	0.9211	0.9392	0.9326	0.8362	0.7649	0.9448	0.9198	<u>0.9098</u>
		ρ (μ_{Mean} , m-MMLU)	0.9342	0.9697	0.9362	0.9689	0.9647	0.9223	0.9406	0.9857	0.9393	<u>0.9513</u>
		ρ (μ_{Max} , m-MMLU)	0.9060	0.9596	0.8946	0.9003	0.8892	0.9386	0.8936	0.9311	0.9565	<u>0.9188</u>
		ρ (μ_{Mean} , m-ARC)	0.9741	0.9706	0.9374	0.9515	0.9562	0.9052	0.9268	0.9693	0.8630	0.9393
		ρ (μ_{Max} , m-ARC)	0.9187	0.9499	0.8736	0.8582	0.8663	0.9297	0.8439	0.9001	0.8298	0.8856
330 331 332	last token	ρ (μ_{Mean} , Belebele)	0.8997	0.9326	0.8491	0.9494	0.9581	0.9141	0.8340	0.9679	0.9467	0.9168
		ρ (μ_{Max} , Belebele)	0.9225	0.9309	0.9127	0.9244	0.9123	0.9125	0.7693	0.9460	0.9218	0.9058
		ρ (μ_{Mean} , m-MMLU)	0.9086	0.9637	0.9370	0.9687	0.9690	0.9771	0.9301	0.9659	0.9700	0.9545
		ρ (μ_{Max} , m-MMLU)	0.8448	0.9297	0.8645	0.9224	0.9177	0.9699	0.8902	0.9161	0.9649	0.9134
		ρ (μ_{Mean} , m-ARC)	0.9190	0.9541	0.9524	0.9536	0.9617	0.9390	0.9146	0.9451	0.7356	<u>0.9195</u>
		ρ (μ_{Max} , m-ARC)	0.8569	0.9147	0.9005	0.8944	0.8879	0.9464	0.8263	0.8859	0.7037	0.8685
333 334 335 336 337	Bible weighted average	ρ (μ_{Mean} , Belebele)	0.8360	0.8530	0.7909	0.8781	0.8974	0.8982	0.8404	0.9118	0.7410	<u>0.8496</u>
		ρ (μ_{Max} , Belebele)	0.8863	0.9001	0.8851	0.9242	0.9302	0.8926	0.8230	0.9337	0.7549	0.8811
		ρ (μ_{Mean} , m-MMLU)	0.8051	0.8886	0.8958	0.9096	0.8964	0.9252	0.9159	0.9093	0.7944	0.8823
		ρ (μ_{Max} , m-MMLU)	0.5501	0.8831	0.7748	0.8683	0.8364	0.9180	0.9085	0.9107	0.7388	<u>0.8210</u>
		ρ (μ_{Mean} , m-ARC)	0.8505	0.8998	0.9188	0.9267	0.9116	0.8940	0.9208	0.9317	0.8623	0.9018
		ρ (μ_{Max} , m-ARC)	0.6070	0.8803	0.8030	0.8769	0.8552	0.8684	0.8879	0.9178	0.8220	<u>0.8354</u>
338 339 340	last token	ρ (μ_{Mean} , Belebele)	0.7656	0.8005	0.5944	0.7934	0.8396	0.9046	0.8299	0.9177	0.8866	0.8147
		ρ (μ_{Max} , Belebele)	0.7844	0.8299	0.5264	0.8000	0.8100	0.9047	0.8048	0.9235	0.8796	0.8070
		ρ (μ_{Mean} , m-MMLU)	0.7194	0.7646	0.6472	0.6068	0.6516	0.8827	0.8692	0.8672	0.8060	0.7572
		ρ (μ_{Max} , m-MMLU)	0.7075	0.6886	0.5037**	0.5228**	0.4461**	0.9079	0.8576	0.8643	0.7994	0.6998
		ρ (μ_{Mean} , m-ARC)	0.7411	0.7754	0.6592	0.5976	0.6494	0.8537	0.8537	0.8927	0.6997	0.7469
		ρ (μ_{Max} , m-ARC)	0.7293	0.7000	0.5190**	0.5335**	0.4853**	0.8494	0.8309	0.8624	0.6867	0.6885

Table 2: Pearson correlation of MEXA using FLORES and Bible data across three tasks. ρ (μ_{Pooling} , Task) is the correlation of MEXA for the corresponding pooling strategy and benchmark. In all settings except **, the p-value is $p < 0.001$. The best average correlations for each task are in **bold**, and the second bests are underlined.

due to Llama 2’s weaker performance on English tasks, it fails to transfer this alignment effectively, leading to comparable non-English performance between Llama 2 and OLMo. This observation is further explored in Section 5.2, where we normalize the expected performance based on the pivot language, namely English.

5.1 MEXA CORRELATION ANALYSIS

We compute sentence embeddings for the selected models using two methods: weighted average based on token positions and last token (see §3.1). We apply mean and max pooling on the MEXA alignment scores across all model layers to derive a single score for each language. In Table 2, we report the correlation between the MEXA scores (computed using both mean- and max-pooling, for the two embedding methods) and task performances. Across all settings, the best overall result (higher correlation) is achieved when embeddings are computed using the **weighted average**, with **mean pooling** as the pooling method. We adopt this configuration as the default setting for MEXA.

FLORES vs Bible. In the default setting, the average Pearson correlation score for the FLORES parallel dataset across different tasks is 0.9300, while for the Bible parallel dataset, it is 0.8779. The reason the Bible scores are generally lower than FLORES is that FLORES data is cleaner and more aligned with modern, standardized texts, whereas the Bible data is older and more specialized. For example for some languages, the orthography of Bible texts no longer matches today’s orthography. In the Bible, Arabic often includes diacritics, which are typically omitted in modern writing and tasks, making the text less familiar to models trained on contemporary data. Additionally, although the Bible dataset has been made parallel, sentence alignment can still be inconsistent due to translation nuances. In contrast, FLORES is carefully curated to ensure high-quality, sentence-level parallelism across languages for machine translation tasks.

Weighted Average vs. Last Token Embeddings. The use of last token embeddings shows promisingly high correlations with the FLORES parallel data; however, for the Bible dataset, the correlation is low in some cases. We believe this may stem from the high occurrence of Bible sentences (especially in English), which leads models to memorize these phrases. Using the WIMBD toolkit (Elazar et al., 2024), we found that, on average, there are 92 times more documents in Dolma 1.7 (Soldaini et al., 2024) containing exact Bible sentences than those in FLORES. Consequently, when using Bible examples, the last token is biased towards predicting the specific memorized next token rather than incorporating context-related signals. Therefore, one should consider the hazard of memorized data when using last token embeddings. The weighted-average method, which takes into account

the influence of multiple tokens, can mitigate the impact of a poor embedding for the last token and enable the model to capture useful information from the other tokens more robustly.

Max Pooling vs. Mean Pooling. In our comparison of mean pooling and max pooling on the Belebele benchmark, we found that mean pooling underestimates low-resource languages (resulting in more MEXA scores near 0), while max pooling correlates better with the Belebele benchmark. This can be explained by the fact that Belebele is an easier task among the three evaluated, allowing even low-resource languages to achieve good scores. Conversely, based on our experiment with m-ARC, max pooling tends to overestimate low-resource languages, making mean pooling more aligned with m-ARC. This can be attributed to m-ARC being the most challenging task among the three, where even medium-resource languages do not achieve high scores. Changing the pooling method from mean to max can be considered when dealing with different levels of understanding.

5.2 DOWNSTREAM PERFORMANCE ESTIMATION

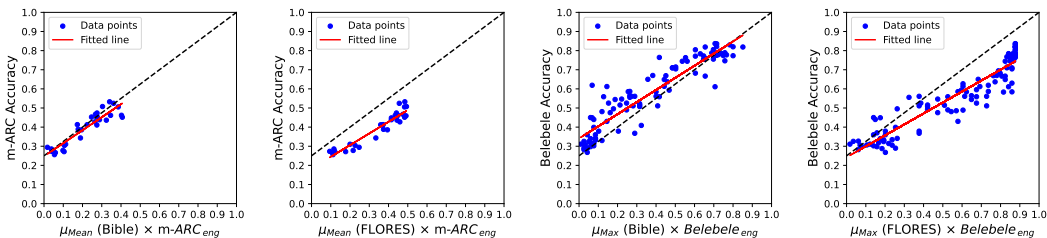


Figure 1: The relationship between MEXA scores of Llama 3.1-8B from the Bible and FLORES, adjusted by the English task performance, for tasks in Belebele and the m-ARC benchmark.

A complete Pearson correlation (i.e. $\rho = 1.0$) indicates that a linear equation perfectly describes the relationship between MEXA and the evaluation benchmarks, with all data points lying on a line. Given the high correlation values shown in Table 2, it is reasonable to conclude that we can fit a line that closely approximates this linear relationship. This line converts the MEXA scores back to downstream task performances. We employed a linear model to predict this line by minimizing the residual sum of squares between the MEXA scores (multiplied by the performance on the English task) and the task performances. We needed to adjust the MEXA scores for this purpose, as the MEXA score for language L_1 indicates how well L_1 is aligned with English but does not reflect the estimated task performance of the model for language L_1 . Of course, this does not change the Pearson correlation, as it is unaffected by linear transformations. The three tasks considered in this paper involve multiple-choice questions with four possible answers for each question, resulting in a chance of being randomly correct of $\frac{1}{4}$. However, the minimum score for MEXA scores is 0. Thus, the ideal slope for the line would be $\frac{3}{4}$ with an intercept of $\frac{1}{4}$ (X-axis: adjusted MEXA scores, Y-axis: task performance). In Figure 1, we plot this relationship for Llama 3.1-8B using the Bible and FLORES parallel datasets for Belebele and m-ARC. We chose mean pooling for Belebele and max pooling for m-ARC, since these pooling methods yield a stronger correlation (see §5.1). The pairs of (slope, intercept) from left to right in the Figure 1 are: (0.6804, 0.2477), (0.6103, 0.1838), (0.6340, 0.3408) and (0.5726, 0.2423). With data points from both high-resource and low-resource languages, this line can be calculated; otherwise, the ideal line may be used as a reference.

Language Coverage. We present the adjusted MEXA score for all languages available in FLORES-200 in Table 5 from Appendix A.5 for the selected models. The languages are categorized into groups ranging from well-covered to not covered. In Table 5, we can clearly see that Llama 3.1-70B and Gemma 2-9B show a higher level of multilinguality than other models.

5.3 MEXA VS ABSOLUTE COSINE SIMILARITY

We compare MEXA with the use of absolute cosine similarities. To begin with, cosine similarity scores are not always directly comparable across models. For example, if a language shows a higher cosine similarity with English for one model than another, it does not necessarily indicate better alignment in the former model. However, MEXA has the advantage of being directly comparable, as

its score does not rely on absolute similarity values. To examine the correlation of both methods with downstream tasks, we conducted the following experiment. We used parallel data from FLORES and downstream task data from the Belebele benchmark, focusing on 116 common labels. For each non-English language, we computed the average absolute cosine similarity for parallel sentences with English, and the average absolute cosine similarity for non-parallel sentences with English. Following the setup by Li et al. (2024f), which employs absolute cosine similarity values to predict the performance and rank of languages, we computed the embeddings using the last-token method and applied mean pooling over layers {5, 10, 15, 20, 25}. We report results using the Gemma 1 and Llama 1 7B models, which are commonly used in our experiments. For a fair comparison, this setup is applied to both absolute cosine similarity and MEXA. For the Gemma 1 model, MEXA achieves a correlation of 0.9260 with downstream task performance, while the absolute cosine similarity for parallel sentences achieves a correlation of 0.7651. Additionally, the correlation between the absolute cosine similarity for parallel and non-parallel sentences is 0.9232. For the Llama 1 model, MEXA achieves a correlation of 0.8365 with downstream task performance, while the absolute cosine similarity for parallel sentences achieves a correlation of 0.6473. Additionally, the correlation between the absolute cosine similarity for parallel and non-parallel sentences is 0.9064. In both models, the absolute cosine similarity method achieved significantly lower correlations with downstream tasks compared to MEXA. This discrepancy arises primarily because, for some languages, the similarity score can be high regardless of whether the sentences are parallel or non-parallel. Furthermore, a low similarity score between two languages does not necessarily indicate weak alignment, as overall similarity scores may be low while parallel sentences still exhibit much higher scores than non-parallel ones.

5.4 VISUALIZATION OF LAYERS

In Figure 2, we show the results of applying MEXA to 20 pairs of language_script from FLORES parallel dataset for Llama 1-7B and Llama 3.1-8B across all 32 layers. We selected these languages from different families, writing systems, and both high- and low-resource categories. The embeddings are computed using weighted average based on token positions. Figure 2 shows that high-resource languages (with more prevalence on the web; see §A.1) achieve higher alignment scores across different layers, while low-resource languages achieve lower scores. In the initial layers, embeddings are more in-language, resulting in lower alignment scores. As embeddings progress to the mid-layers, they become more aligned with the main language of the LLM, i.e., English.

MEXA is comparable between models as long as the same parallel dataset and setting is used to obtain the MEXA scores. Figure 2 shows that in many languages, particularly high-resource languages, Llama 3.1 achieves a significantly higher alignment score than its predecessor, Llama 1. Although Llama 3.1 exhibits better alignment scores with English for medium and low-resource languages, there is still significant room for improvement. Comparing Arabic (arb_Arab) with its romanized version (arb_Latn), we see that both Llama 1 and Llama 3.1 models perform better in the native script than in the Latin script, even though Llama 1’s tokenizer for Arabic is essentially a character-based tokenizer. In general, for very low-resource languages, those in Latin script tend to have higher alignment scores, likely because the tokenization is more favorable for Latin characters.

In Figure 3, we display the t-SNE (Van der Maaten & Hinton, 2008) plots of the embeddings of Figure 2 from 3 different layers of Llama 3.1: embedding layer 0, mid-layer 13, and last layer 32. We assign a different color to each language. For layers 0 and 32, the embeddings are more language-specific, while in the mid-layer, they become more language-neutral. Languages that maintain their language-specific embeddings in the mid-layer are clustered separately and, notably, correspond to the very low-resource languages that receive the worst alignment scores from MEXA.

6 CONCLUSION

We introduce MEXA, a method for assessing the multilingual capabilities of English-centric large language models (LLMs). MEXA builds on the observation that English-centric LLMs semantically use English as a pivot language in their intermediate layers. MEXA computes the alignment between non-English languages and English using parallel sentences, estimating the transfer of language understanding capabilities from English to other languages through this alignment. This metric can be useful in estimating task performance, provided we know the English performance in the task and the

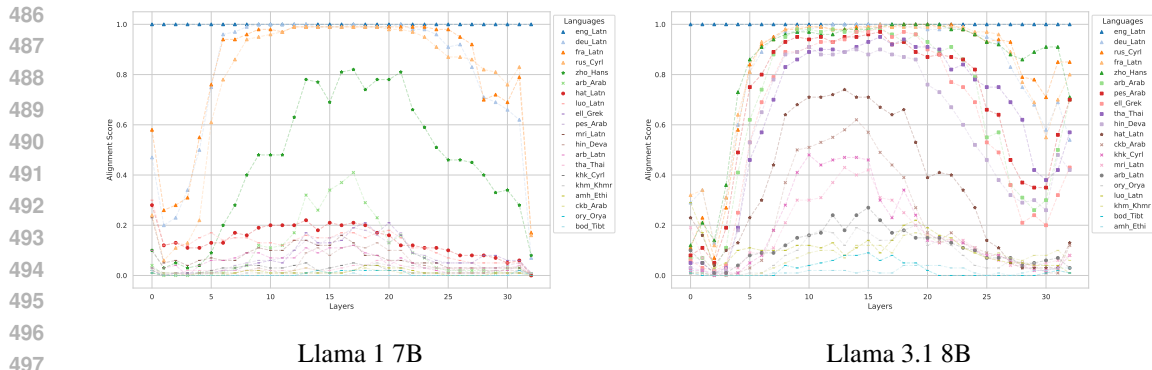


Figure 2: Llama 1 vs. Llama 3.1 agreement score for different languages across all layers. Best performance markers in order: \triangle , \square , \star , \times , \circ , $-$

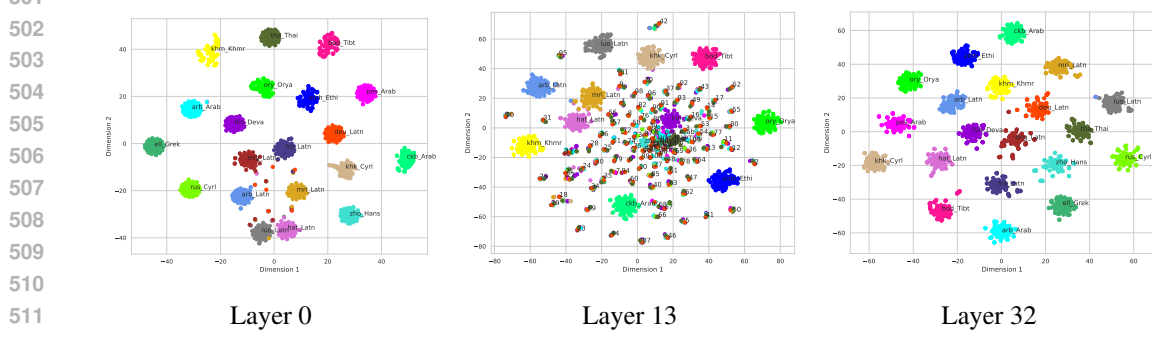


Figure 3: Llama 3.1 t-SNE plots for 3 different layers. As shown, in the mid-layers, the embeddings become more language-neutral. The numbers shown in the mid-layers are the IDs of English sentences that are scattered.

alignment score between languages derived from a parallel dataset. Through different studies with two parallel datasets (FLORES-200 and the Bible); a diverse range of LLMs including the Llama family, Gemma family, Mistral, and OLMo, and three downstream tasks (Belebele, m-MMLU, and m-ARC), we demonstrated that MEXA provides a reliable estimation of multilingual performance. For MEXA score calculations, multiple design analyses are conducted to explore the impact of token-level pooling for embeddings and layer-level pooling in computing alignment scores. While MEXA shows high correlation across most configurations, a weighted average of tokens combined with mean pooling delivers the best results. The results reveal a promising average Pearson correlation of 0.90 with established downstream tasks across nine models and two parallel dataset. Overall, MEXA proves to be a valuable method for practitioners aiming to assess the multilingual capabilities of English-centric LLMs, paving the way for future efforts to expand these models to a wider range of underrepresented languages.

7 LIMITATION

MEXA provides a rough estimate of the multilingual capabilities of pre-trained English-centric LLMs. Different tasks offer diverse perspectives on the abilities of LLMs, and MEXA cannot replace all of them. Our goal is to highlight the multilingual potential of English-centric LLMs and propose a simple way to evaluate them. We hope this encourages the development of more multilingual LLMs, even though they are likely to contain large shares of English data. Additionally, it is important to note that answers across languages do not always need to be fully aligned (Naous et al., 2024), and for such cases, language- and culture-specific evaluation benchmarks should be developed.

REFERENCES

- 540
541
542 David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi,
543 Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka
544 Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon
545 Kabongo, Foutse Yuehghoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither
546 Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge,
547 and Pontus Stenetorp. Irokobench: A new benchmark for African languages in the age of large
548 language models, 2024. URL <https://arxiv.org/abs/2406.03368>.
- 549 Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. IndicXNLI: Evaluating multilingual
550 inference for Indian languages. In *Proceedings of the 2022 Conference on Empirical Methods in
551 Natural Language Processing*, pp. 10994–11006, Abu Dhabi, United Arab Emirates, December
552 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.755. URL
553 <https://aclanthology.org/2022.emnlp-main.755>.
- 554 Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain,
555 Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana
556 Sitaram. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Confer-
557 ence on Empirical Methods in Natural Language Processing*, pp. 4232–4267, Singapore, Decem-
558 ber 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.258.
559 URL <https://aclanthology.org/2023.emnlp-main.258>.
- 560 Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent
561 Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram.
562 MEGEVERSE: Benchmarking large language models across languages, modalities, models and
563 tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Asso-
564 ciation for Computational Linguistics: Human Language Technologies (Volume 1: Long Pa-
565 pers)*, pp. 2598–2637, Mexico City, Mexico, June 2024. Association for Computational Linguis-
566 tics. doi: 10.18653/v1/2024.naacl-long.143. URL <https://aclanthology.org/2024.naacl-long.143>.
- 567
568 Anthropic. The Claude 3 model family: Opus, sonnet, haiku, 2023. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.
- 569
570
571 Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of mono-
572 lingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Com-
573 putational Linguistics*, pp. 4623–4637, 2020.
- 574
575 Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald
576 Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The Bele-
577 bele benchmark: a parallel reading comprehension dataset in 122 language variants, 2024. URL
578 <https://arxiv.org/abs/2308.16884>.
- 579
580 Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Al-
581 ham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi,
582 Julen Etxaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa
583 Jaiswal, Wilson Y. Lee, Haonan Li, Charles Lovering, Niklas Muennighoff, Ellie Pavlick, Ja-
584 son Phang, Aviya Skowron, Samson Tan, Xiangru Tang, Kevin A. Wang, Genta Indra Winata,
585 François Yvon, and Andy Zou. Lessons from the trenches on reproducible evaluation of language
586 models, 2024. URL <https://arxiv.org/abs/2405.14782>.
- 587
588 BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić,
589 Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé,
590 et al. BLOOM: A 176b-parameter open-access multilingual language model, 2023. URL
591 <https://arxiv.org/abs/2211.05100>.
- 592
593 Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with
subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146,
2017. doi: 10.1162/tacl.a.00051. URL <https://aclanthology.org/Q17-1010>.

- 594 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
595 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are
596 few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020a.
597
- 598 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
599 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. GPT-3 dataset lan-
600 guage statistics, 2020b. URL [https://github.com/openai/gpt-3/tree/master/
601 dataset_statistics](https://github.com/openai/gpt-3/tree/master/dataset_statistics).
- 602 Steven Cao, Nikita Kitaev, and Dan Klein. Multilingual alignment of contextual word repre-
603 sentations. In *International Conference on Learning Representations*, 2020. URL [https://
604 //openreview.net/forum?id=r1xCMYBtPS](https://openreview.net/forum?id=r1xCMYBtPS).
- 605 Grzegorz Chrupała and Afra Alishahi. Correlating neural and symbolic representations of language.
606 In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp.
607 2952–2962, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/
608 v1/P19-1283. URL <https://aclanthology.org/P19-1283>.
- 609 Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev,
610 and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in
611 typologically diverse languages. *Transactions of the Association for Computational Linguistics*,
612 8:454–470, 2020.
613
- 614 Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and
615 Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning chal-
616 lenge, 2018. URL <https://arxiv.org/abs/1803.05457>.
- 617 Common Crawl. Statistics of common crawl monthly archives, 2024. URL [https://
618 commoncrawl.github.io/cc-crawl-statistics/plots/languages.html](https://commoncrawl.github.io/cc-crawl-statistics/plots/languages.html).
- 619 Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *Advances
620 in Neural Information Processing Systems*, volume 32, pp. 7059–7069. Curran Associates, Inc.,
621 2019.
622
- 623 Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger
624 Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In
625 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp.
626 2475–2485, Brussels, Belgium, October–November 2018. Association for Computational Lin-
627 guistics. doi: 10.18653/v1/D18-1269. URL <https://aclanthology.org/D18-1269>.
- 628 Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek,
629 Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Un-
630 supervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual
631 Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020a.
632 Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL [https://
633 aclanthology.org/2020.acl-main.747](https://aclanthology.org/2020.acl-main.747).
- 634 Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-
635 lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting
636 of the Association for Computational Linguistics*, pp. 6022–6034, Online, July 2020b. Asso-
637 ciation for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.536. URL [https://
638 aclanthology.org/2020.acl-main.536](https://aclanthology.org/2020.acl-main.536).
- 639 Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Uni-
640 versal Dependencies. *Computational Linguistics*, 47(2):255–308, June 2021. doi: 10.1162/
641 coli.a.00402. URL <https://aclanthology.org/2021.cl-2.11>.
- 642
- 643 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep
644 bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of
645 the North American Chapter of the Association for Computational Linguistics: Human Language
646 Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June
647 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL [https://
//aclanthology.org/N19-1423](https://aclanthology.org/N19-1423).

- 648 Sumanth Doddapaneni, Rahul Aralikkatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra,
649 Anoop Kunchukuttan, and Pratyush Kumar. Towards leaving no Indic language behind: Building
650 monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the*
651 *61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12402–12426, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.693. URL <https://aclanthology.org/2023.acl-long.693>.
- 655 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
656 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models,
657 2024. URL <https://arxiv.org/abs/2407.21783>.
- 659 Pavel Efimov, Leonid Boytsov, Elena Arslanova, and Pavel Braslavski. The impact of cross-
660 lingual adjustment of contextual word representations on zero-shot transfer. In *Advances in*
661 *Information Retrieval*, pp. 51–67, Cham, 2023. Springer Nature Switzerland. URL https://link.springer.com/10.1007/978-3-031-28241-6_4.
- 663 Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk,
664 Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi,
665 Noah A. Smith, and Jesse Dodge. What’s in my big data? In *The Twelfth International*
666 *Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=RvfPnOkPV4>.
- 668 Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the ge-
669 ometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference*
670 *on Empirical Methods in Natural Language Processing and the 9th International Joint Conference*
671 *on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China,
672 November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL
673 <https://aclanthology.org/D19-1006>.
- 674 Christian Federmann, Tom Kocmi, and Ying Xin. NTREX-128 – news test references for MT
675 evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual*
676 *Evaluation*, pp. 21–24, Online, November 2022. Association for Computational Linguistics. URL
677 <https://aclanthology.org/2022.sumeval-1.4>.
- 678 Jay Gala, Pranjal A Chitale, AK Raghavan, Varun Gumma, Sumanth Doddapaneni, Janki Atul
679 Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, et al. Indic-
680 trans2: Towards high-quality and accessible machine translation models for all 22 scheduled
681 indian languages. *Transactions on Machine Learning Research*.
- 683 Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster,
684 Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muen-
685 nighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang
686 Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-
687 shot language model evaluation, December 2023. URL <https://zenodo.org/records/10256836>.
- 688 Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu,
689 Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly
690 capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- 693 Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya
694 Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open
695 models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024a.
- 696 Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhu-
697 patiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma
698 2: Improving open language models at a practical size, 2024b. URL <https://arxiv.org/abs/2408.00118>.
- 700 Omid Ghahroodi, Marzia Nouri, Mohammad Vali Sanian, Alireza Sahebi, Doratossadat Dastgheib,
701 Ehsaneddin Asgari, Mahdieh Soleymani Baghshah, and Mohammad Hossein Rohban. Khayyam

- 702 challenge (PersianMMLU): Is your LLM truly wise to the Persian language?, 2024. URL
703 <https://arxiv.org/abs/2404.06644>.
704
- 705 Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord,
706 Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. OLMo: Accelerat-
707 ing the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.
- 708 Katharina Hämmerl, Alina Fastowski, Jindřich Libovický, and Alexander Fraser. Exploring
709 anisotropy and outliers in multilingual language models for cross-lingual semantic sentence sim-
710 ilarity. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7023–7037,
711 Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.
712 findings-acl.439. URL <https://aclanthology.org/2023.findings-acl.439>.
- 713 Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. Understanding cross-lingual
714 Alignment—A survey. In *Findings of the Association for Computational Linguistics ACL*
715 *2024*, pp. 10922–10943, Bangkok, Thailand and virtual meeting, August 2024. Association
716 for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.649. URL [https://](https://aclanthology.org/2024.findings-acl.649)
717 aclanthology.org/2024.findings-acl.649.
- 718 Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin
719 Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive
720 summarization for 44 languages. In *Findings of the Association for Computational Linguistics:*
721 *ACL-IJCNLP 2021*, pp. 4693–4703, Online, August 2021. Association for Computational Lin-
722 guistics. doi: 10.18653/v1/2021.findings-acl.413. URL [https://aclanthology.org/](https://aclanthology.org/2021.findings-acl.413)
723 [2021.findings-acl.413](https://aclanthology.org/2021.findings-acl.413).
- 724 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Ja-
725 cob Steinhardt. Measuring massive multitask language understanding, 2021. URL [https:](https://arxiv.org/abs/2009.03300)
726 [//arxiv.org/abs/2009.03300](https://arxiv.org/abs/2009.03300).
727
- 728 Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson.
729 XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual general-
730 isation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119
731 of *Proceedings of Machine Learning Research*, pp. 4411–4421. PMLR, 13–18 Jul 2020. URL
732 <https://proceedings.mlr.press/v119/hu20b.html>.
- 733 Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou.
734 Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In
735 *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and*
736 *the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp.
737 2485–2494, Hong Kong, China, November 2019. Association for Computational Linguistics. doi:
738 10.18653/v1/D19-1252. URL <https://aclanthology.org/D19-1252>.
- 739 Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot,
740 Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al.
741 Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023a.
742
- 743 Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bam-
744 ford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al.
745 Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.
- 746 Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence
747 embeddings with large language models. *arXiv preprint arXiv:2307.16645*, 2023b.
748
- 749 Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and
750 fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual*
751 *Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online, July 2020.
752 Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL [https:](https://aclanthology.org/2020.acl-main.560)
753 [//aclanthology.org/2020.acl-main.560](https://aclanthology.org/2020.acl-main.560).
- 754 Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. GLoLID: Lan-
755 guage identification for low-resource languages. In *Findings of the Association for Compu-*
tational Linguistics: EMNLP 2023, pp. 6155–6218, Singapore, December 2023. Association

- 756 for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.410. URL <https://aclanthology.org/2023.findings-emnlp.410>.
- 757
758
- 759 Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. GlotScript: A resource and tool
760 for low resource writing system identification. In *Proceedings of the 2024 Joint Interna-*
761 *tional Conference on Computational Linguistics, Language Resources and Evaluation (LREC-*
762 *COLING 2024)*, pp. 7774–7784, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.687>.
- 763
- 764 Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of*
765 *Machine Translation Summit X: Papers*, pp. 79–86, Phuket, Thailand, September 13-15 2005.
766 URL <https://aclanthology.org/2005.mtsummit-papers.11>.
- 767
- 768 Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural net-
769 work representations revisited. In *Proceedings of the 36th International Conference on Machine*
770 *Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–
771 15 Jun 2019. URL <https://proceedings.mlr.press/v97/kornblith19a.html>.
- 772
- 773 Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi,
774 Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov,
775 and Timothy Baldwin. ArabicMMLU: Assessing massive multitask language understanding in
776 arabic, 2024. URL <https://arxiv.org/abs/2402.12840>.
- 777
- 778 Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. Cross-lingual align-
779 ment methods for multilingual BERT: A comparative study. In *Findings of the Association*
780 *for Computational Linguistics: EMNLP 2020*, pp. 933–942, Online, November 2020. Associ-
781 ation for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.83. URL <https://aclanthology.org/2020.findings-emnlp.83>.
- 782
- 783 Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. WikiLingua: A new
784 benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Associa-*
785 *tion for Computational Linguistics: EMNLP 2020*, pp. 4034–4048, Online, November 2020.
786 Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.360. URL
787 <https://aclanthology.org/2020.findings-emnlp.360>.
- 788
- 789 Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and
790 Thien Nguyen. ChatGPT beyond English: Towards a comprehensive evaluation of large language
791 models in multilingual learning. In *Findings of the Association for Computational Linguistics:*
792 *EMNLP 2023*, pp. 13171–13189, Singapore, December 2023a. Association for Computational
793 Linguistics. doi: 10.18653/v1/2023.findings-emnlp.878. URL <https://aclanthology.org/2023.findings-emnlp.878>.
- 794
- 795 Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien
796 Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforc-
797 e-ment learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Meth-*
798 *ods in Natural Language Processing: System Demonstrations*, pp. 318–327, Singapore, Decem-
799 ber 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.28.
800 URL <https://aclanthology.org/2023.emnlp-demo.28>.
- 801
- 802 Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evalu-
803 ating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of*
804 *the Association for Computational Linguistics*, pp. 7315–7330, 2020.
- 805
- 806 Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. Improving in-context learning of
807 multilingual generative language models with cross-lingual alignment. In *Proceedings of the 2024*
808 *Conference of the North American Chapter of the Association for Computational Linguistics:*
809 *Human Language Technologies (Volume 1: Long Papers)*, pp. 8058–8076, Mexico City, Mexico,
810 June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.445.
811 URL <https://aclanthology.org/2024.naacl-long.445>.
- 812
- 813 Daoyang Li, Mingyu Jin, Qingcheng Zeng, Haiyan Zhao, and Mengnan Du. Exploring multilingual
814 probing in large language models: A cross-language analysis, 2024b. URL <https://arxiv.org/abs/2409.14459>.

- 810 Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy
811 Baldwin. CMMLU: Measuring massive multitask language understanding in Chinese, 2024c.
812 URL <https://arxiv.org/abs/2306.09212>.
- 813
- 814 Jiahuan Li, Shujian Huang, Xinyu Dai, and Jiajun Chen. PreAlign: Boosting cross-lingual transfer
815 by early establishment of multilingual alignment, 2024d. URL <https://arxiv.org/abs/2407.16222>.
- 816
- 817 Xianming Li and Jing Li. AoE: Angle-optimized embeddings for semantic textual similarity. In *Pro-*
818 *ceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1:*
819 *Long Papers)*, pp. 1825–1839, Bangkok, Thailand, August 2024. Association for Computational
820 Linguistics. URL <https://aclanthology.org/2024.acl-long.101>.
- 821
- 822 Xiaochen Li, Zheng-Xin Yong, and Stephen H Bach. Preference tuning for toxicity mitigation
823 generalizes across languages. *arXiv preprint arXiv:2406.16235*, 2024e.
- 824
- 825 Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. Quan-
826 tifying multilingual performance of large language models across languages, 2024f. URL
827 <https://arxiv.org/abs/2404.11553>.
- 828
- 829 Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle
830 Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh
831 Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva,
832 Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative
833 language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natu-*
834 *ral Language Processing*, pp. 9019–9052, Abu Dhabi, United Arab Emirates, December 2022.
835 Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.616. URL
836 <https://aclanthology.org/2022.emnlp-main.616>.
- 837
- 838 Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie
839 and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources*
840 *and Evaluation (LREC’16)*, pp. 923–929, Portorož, Slovenia, May 2016. European Language
841 Resources Association (ELRA). URL <https://aclanthology.org/L16-1147>.
- 842
- 843 Weihao Liu, Ning Wu, Wenbiao Ding, Shining Liang, Ming Gong, and Dongmei Zhang. Towards
844 truthful multilingual large language models: Benchmarking and alignment strategies, 2024a. URL
845 <https://arxiv.org/abs/2406.14434>.
- 846
- 847 Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Ayyoob Imani, Orgest Xhelili, Haotian Ye,
848 Chunlan Ma, François Yvon, and Hinrich Schütze. How transliterations improve crosslingual
849 alignment, 2024b. URL <https://arxiv.org/abs/2409.17326>.
- 850
- 851 Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning llama for multi-stage
852 text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and*
853 *Development in Information Retrieval*, pp. 2421–2425, 2024.
- 854
- 855 Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and
856 Sebastian Ruder. How does quantization affect multilingual LLMs?, 2024. URL <https://arxiv.org/abs/2407.03211>.
- 857
- 858 Thomas Mayer and Michael Cysouw. Creating a massively parallel Bible corpus. In *Proceed-*
859 *ings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*,
860 pp. 3158–3163, Reykjavik, Iceland, May 2014. European Language Resources Association
861 (ELRA). URL [http://www.lrec-conf.org/proceedings/lrec2014/pdf/220_](http://www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf)
862 [Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf).
- 863
- 864 Meta. Llama 3, 2024. URL <https://llama.meta.com/llama3/>.
- 865
- 866 Niklas Muennighoff. SGPT: GPT sentence embeddings for semantic search. *arXiv preprint*
867 *arXiv:2202.08904*, 2022.

- 864 Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven
865 Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang,
866 Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Al-
867 bert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask
868 finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational*
869 *Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, Toronto, Canada, July 2023. Asso-
870 ciation for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL <https://aclanthology.org/2023.acl-long.891>.
871
- 872 Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. First align, then predict: Under-
873 standing the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference*
874 *of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp.
875 2214–2231, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/
876 2021.eacl-main.189. URL <https://aclanthology.org/2021.eacl-main.189>.
877
- 878 Tarek Naous, Michael Ryan, Alan Ritter, and Wei Xu. Having beer after prayer? measuring
879 cultural bias in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar
880 (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Lin-*
881 *guistics (Volume 1: Long Papers)*, pp. 16366–16393, Bangkok, Thailand, August 2024. As-
882 sociation for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.862. URL <https://aclanthology.org/2024.acl-long.862>.
883
- 884 Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming
885 Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris
886 Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski
887 Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter
888 Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training, 2022. URL
889 <https://arxiv.org/abs/2201.10005>.
- 890 NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield,
891 Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler
892 Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez,
893 Prangthip Hansanti, John Hoffman, Searley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shan-
894 non Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela
895 Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko,
896 Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left be-
897 hind: Scaling human-centered machine translation, 2022. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2207.04672)
898 [2207.04672](https://arxiv.org/abs/2207.04672).
- 899 Odunayo Ogundepo, Tajuddeen Gwadabe, Clara Rivera, Jonathan Clark, Sebastian Ruder, David
900 Adelani, Bonaventure Dossou, Abdou Diop, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba,
901 Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Emezue, Albert Kahira, Shamsuddeen
902 Muhammad, Akintunde Oladipo, Abraham Owodunni, Atnafu Tonja, Iyanuoluwa Shode, Akari
903 Asai, Anuoluwapo Aremu, Ayodele Awokoya, Bernard Opoku, Chiamaka Chukwuneke, Chris-
904 tine Mwase, Clemencia Siro, Stephen Arthur, Tunde Ajayi, Verrah Otiende, Andre Rubungo,
905 Boyd Sinkala, Daniel Ajisafe, Emeka Onwuegbuzia, Falalu Lawan, Ibrahim Ahmad, Jesujoba
906 Alabi, Chinedu Mbonu, Mofetoluwa Adeyemi, Mofya Phiri, Orevaoghene Ahia, Ruqayya Iro,
907 and Sonia Adhiambo. Cross-lingual open-retrieval question answering for African languages.
908 In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14957–14972,
909 Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.
910 findings-emnlp.997. URL [https://aclanthology.org/2023.findings-emnlp.](https://aclanthology.org/2023.findings-emnlp.997)
911 [997](https://aclanthology.org/2023.findings-emnlp.997).
- 912 OpenAI. Introducing ChatGPT, 2022. URL <https://openai.com/blog/chatgpt>.
- 913 OpenAI. Multilingual massive multitask language understanding (MMMLU), 2024. URL <https://huggingface.co/datasets/openai/MMMLU>.
914 <https://huggingface.co/datasets/openai/MMMLU>.
915
- 916 OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Floren-
917 cia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al.
GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

- 918 Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-
919 lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting*
920 *of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1946–1958, 2017.
921
- 922 Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen.
923 XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020*
924 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2362–2376,
925 Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.
926 emnlp-main.185. URL <https://aclanthology.org/2020.emnlp-main.185>.
- 927 Sara Rajaei and Mohammad Taher Pilehvar. An isotropy analysis in the multilingual BERT embed-
928 ding space. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1309–
929 1316, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/
930 v1/2022.findings-acl.103. URL [https://aclanthology.org/2022.findings-acl.](https://aclanthology.org/2022.findings-acl.103)
931 103.
- 932 Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual us-
933 ing knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in*
934 *Natural Language Processing (EMNLP)*, pp. 4512–4525, Online, November 2020. Associa-
935 tion for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.365. URL <https://aclanthology.org/2020.emnlp-main.365>.
936
- 937 Anton Schäfer, Shauli Ravfogel, Thomas Hofmann, Tiago Pimentel, and Imanol Schlag. The role of
938 language imbalance in cross-lingual generalisation: Insights from cloned language experiments,
939 2024. URL <https://arxiv.org/abs/2404.07982>.
- 940
- 941 Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin
942 Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith
943 Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński,
944 Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai,
945 Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrman,
946 Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara
947 Hooker. Aya dataset: An open-access collection for multilingual instruction tuning, 2024.
- 948 Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur,
949 Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh
950 Jha, Sachin Kumar, Li Lucy, Xinxu Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas
951 Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle
952 Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke
953 Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and
954 Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research,
955 2024. URL <https://arxiv.org/abs/2402.00159>.
- 956 Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi,
957 Cheonbok Park, Kang Min Yoo, and Stella Biderman. KMMLU: Measuring massive multitask
958 language understanding in Korean, 2024. URL <https://arxiv.org/abs/2402.11548>.
- 959 Stanford CRFM. Holistic evaluation of language models - MMLU leaderboard, 2024. URL <https://crfm.stanford.edu/helm/mmlu/latest/#/leaderboard>.
- 960
- 961 Tatoeba Community. Tatoeba collection, 2006. URL [https://tatoeba.org/en/](https://tatoeba.org/en/downloads)
962 [downloads](https://tatoeba.org/en/downloads).
- 963
- 964 Atula Tejaswi, Nilesh Gupta, and Eunsol Choi. Exploring design choices for building language-
965 specific LLMs. *arXiv preprint arXiv:2406.14670*, 2024.
- 966
- 967 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée
968 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and
969 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.
- 970
- 971 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko-
lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open founda-
tion and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

- 972 Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine*
973 *learning research*, 9(11), 2008.
- 974
- 975 W3Techs. Usage statistics of content languages for websites, 2024. URL [https://w3techs.](https://w3techs.com/technologies/overview/content_language)
976 [com/technologies/overview/content_language](https://w3techs.com/technologies/overview/content_language).
- 977
- 978 Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improv-
979 ing text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting*
980 *of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916,
981 Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/
982 2024.acl-long.642. URL <https://aclanthology.org/2024.acl-long.642>.
- 983 Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do Llamas work in En-
984 glish? on the latent language of multilingual transformers, 2024. URL [https://arxiv.org/
985 abs/2402.10588](https://arxiv.org/abs/2402.10588).
- 986
- 987 Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. PAWS-X: A cross-lingual adversarial
988 dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical*
989 *Methods in Natural Language Processing and the 9th International Joint Conference on*
990 *Natural Language Processing (EMNLP-IJCNLP)*, pp. 3687–3692, Hong Kong, China, November
991 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1382. URL
992 <https://aclanthology.org/D19-1382>.
- 993 Jiacheng Ye, Xijia Tao, and Lingpeng Kong. Language versatilists vs. specialists: An empirical
994 revisiting on multilingual transfer ability, 2023. URL [https://arxiv.org/abs/2306.](https://arxiv.org/abs/2306.06688)
995 [06688](https://arxiv.org/abs/2306.06688).
- 996
- 997 Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Şenel, Anna Korhonen, and Hinrich Schütze. Turkish-
998 MMLU: Measuring massive multitask language understanding in Turkish, 2024. URL <https://arxiv.org/abs/2407.12402>.
- 999
- 1000 Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilin-
1001 gual neural machine translation and zero-shot translation. In *Proceedings of the 58th An-*
1002 *ual Meeting of the Association for Computational Linguistics*, pp. 1628–1639, Online, July
1003 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.148. URL
1004 <https://aclanthology.org/2020.acl-main.148>.
- 1005
- 1006 Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3Exam:
1007 A multilingual, multimodal, multilevel benchmark for examining large language models. In
1008 *Advances in Neural Information Processing Systems*, volume 36, pp. 5484–5505. Curran
1009 Associates, Inc., 2023a. URL [https://proceedings.neurips.cc/paper_files/
1010 paper/2023/file/117c5c8622b0d539f74f6d1fb082a2e9-Paper-Datasets_
1011 and_Benchmarks.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/117c5c8622b0d539f74f6d1fb082a2e9-Paper-Datasets_and_Benchmarks.pdf).
- 1012
- 1013 Zhen-Ru Zhang, Chuanqi Tan, Songfang Huang, and Fei Huang. VECO 2.0: Cross-lingual language
1014 model pre-training with multi-granularity contrastive learning, 2023b. URL [https://arxiv.](https://arxiv.org/abs/2304.08205)
1015 [org/abs/2304.08205](https://arxiv.org/abs/2304.08205).
- 1016
- 1017 Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large
1018 language models handle multilingualism?, 2024. URL [https://arxiv.org/abs/2402.](https://arxiv.org/abs/2402.18815)
1019 [18815](https://arxiv.org/abs/2402.18815).
- 1020
- 1021 Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Mu-
1022 rawaki, and Sadao Kurohashi. Beyond english-centric LLMs: What language do multilingual
1023 language models think in?, 2024. URL <https://arxiv.org/abs/2408.10811>.
- 1024
- 1025 Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. Ques-
1026 tion translation training for better multilingual reasoning, 2024. URL [https://arxiv.org/
1027 abs/2401.07817](https://arxiv.org/abs/2401.07817).

A APPENDIX

A.1 DISTRIBUTION OF PRE-TRAINING DATA IN LLMs

The distribution of languages in the training data of state-of-the-art LLMs is rarely fully documented. Llama 2 (Touvron et al., 2023b) is a counter-example and its authors have disclosed the language distribution use in pretraining. Their analysis uses the FastText (Bojanowski et al., 2017) language identification tool and a threshold of 0.5 for the language detection. We reproduce Touvron et al. (2023b, Table 10), which lists 27 languages with percentages greater than 0.005% in the Llama 2 pre-training data, in Table 3. English, with 89.70%, constitutes the vast majority of the training data.

All the languages listed in Table 3 have a presence of more than 0.10% (top 35 languages) on the web according to the W3Techs report (W3Techs, 2024) or more than 0.15% (top 36 languages) according to CommonCrawl (first three snapshots of 2024) (Common Crawl, 2024). However, not all of the most prevalent languages on the web appear in Table 3. The following 9 languages are missing, most of which use non-Latin writing systems: Turkish (tur_Latn), Persian (pes_Arab), Arabic (ara_Arab), Greek (ell_Grek), Hebrew (heb_Hebr), Thai (tha_Thai), Hindi (hin_Deva), Slovak (slk_Latn), and Lithuanian (lit_Latn).

The distribution of data in the training of English-centric LLMs is not the same as on the web, but it does have some correlation. The amount of English in LLM pre-training data is significantly larger than for other languages. This is also observable for GPT-3 (Brown et al., 2020a), where more than 92% of the training texts was in English (Brown et al., 2020b). The rest of the top languages in the data of such models are mostly high-resource languages, which have the most available data on the web (top 36 languages). However, in some models, this could be adjusted by design, for example, to make writing systems with non-Latin languages less prominent (as seen in Llama 2). This weakens the correlation between LLMs’ pretraining data and the web.

Language	Common Script	Percent	Language	Common Script	Percent
English (eng)	Latn	89.70%	Ukrainian (ukr)	Cyrl	0.07%
Unknown (unk)	-	8.38%	Korean (kor)	Hang	0.06%
German (deu)	Latn	0.17%	Catalan (cat)	Latn	0.04%
French (fra)	Latn	0.16%	Serbian (srp)	Cyrl/Latn	0.04%
Swedish (swe)	Latn	0.15%	Indonesian (ind)	Latn	0.03%
Chinese (zho)	Hans/Hant	0.13%	Czech (ces)	Latn	0.03%
Spanish (spa)	Latn	0.13%	Finnish (fin)	Latn	0.03%
Russian (rus)	Cyrl	0.13%	Hungarian (hun)	Latn	0.03%
Dutch (nld)	Latn	0.12%	Norwegian (nor)	Latn	0.03%
Italian (ita)	Latn	0.11%	Romanian (ron)	Latn	0.03%
Japanese (jpn)	Jpan	0.10%	Bulgarian (bul)	Cyrl	0.02%
Polish (pol)	Latn	0.09%	Danish (dan)	Latn	0.02%
Portuguese (por)	Latn	0.09%	Slovenian (slv)	Latn	0.01%
Vietnamese (vie)	Latn	0.08%	Croatian (hrv)	Latn	0.01%

Table 3: Language distribution in the pretraining data for Llama 2. The large “Unknown” category is partially composed of programming code data. Common scripts are sourced from the GlotScript resource (Kargaran et al., 2024).

A.2 MULTILINGUAL EVALUATION BENCHMARKS

Multilingual evaluation methods and the development of benchmarks not only facilitate the assessment of diverse language representations in LLMs but also help in monitoring cross-lingual generalization, to assess the effect of quantization across multiple languages (Marchisio et al., 2024), the development of language-specific models (Tejaswi et al., 2024), and the optimization of safety preferences (Li et al., 2024e), among others. In Table 4, we list benchmarks with the largest language coverage. This list includes benchmarks referenced by MEGA (Ahuja et al., 2023), MEGA-VERSE (Ahuja et al., 2024), xP3 (Muennighoff et al., 2023), the Aya collection (Singh et al., 2024), the lm-evaluation-harness framework (Gao et al., 2023; Biderman et al., 2024), and inter alia. These datasets comprise a mix of translated datasets, some human-translated or verified by native speakers such as AfriXNLI (Adelani et al., 2024) and some relying only on machine translation Lai et al. (2023b). Additionally, there are datasets created independently for each language, such as XLSum (Hasan et al., 2021), where the data is not parallel and the size of the data varies between languages.

Despite the efforts reflected in Table 4, the community is still lacking highly multilingual benchmarks for tasks such as natural language understanding or text generation.

Dataset	Task	# Languages
XNLI (Conneau et al., 2018)	Natural Language Inference	15
IndicXNLI (Aggarwal et al., 2022)	Natural Language Inference	11
AfriXNLI (Adelani et al., 2024)	Natural Language Inference	15
m_HellaSwag (Lai et al., 2023b)	Natural Language Inference	31
PAWS-X (Yang et al., 2019)	Paraphrase Identification	7
XCOPA (Ponti et al., 2020)	Commonsense Reasoning	11
XStoryCloze (Lin et al., 2022)	Commonsense Reasoning	11
m-ARC (Lai et al., 2023b)	Common Sense Reasoning	31
TyDiQA (Clark et al., 2020)	Question Answering	11
MLQA (Lewis et al., 2020)	Question Answering	7
XQuAD (Artetxe et al., 2020)	Question Answering	11
IndicQA (Doddapaneni et al., 2023)	Question Answering	10
AfriQA (Ogundepo et al., 2023)	Question Answering	10
m_TruthfulQA (Lai et al., 2023b)	Multiple Choice Question Answering	31
UDPOS 2.7 (de Marneffe et al., 2021)	Part of Speech Tagging	104
WikiANN (Pan et al., 2017)	Name Entity Recognition	282
XLSum (Hasan et al., 2021)	Summarization	44
WikiLingua (Ladhak et al., 2020)	Summarization	18
Belebele (Bandarkar et al., 2024)	Multiple Choice Reading Comprehension	115
AfriMMLU (Adelani et al., 2024)	Multiple Choice Knowledge Question Answering	17
m-MMLU (Lai et al., 2023b)	Multiple Choice Knowledge Question Answering	31
MMMLU (OpenAI, 2024)	Multiple Choice Knowledge Question Answering	15
M3Exam (Zhang et al., 2023a)	Multimodal Multiple Choice Knowledge Question Answering	9

Table 4: Multilingual evaluation benchmarks

A.3 SEMANTIC SIMILARITY IN MULTILINGUAL EMBEDDINGS

There are other ways to compute similarity between languages, such as Representational Similarity Analysis (RSA) (Chrupała & Alishahi, 2019) and Central Kernel Alignment (CKA) (Kornblith et al., 2019). RSA involves first computing the cosine similarity for sentence embeddings within each language, then correlating these in-language similarities with those in other languages. CKA, another metric, is adopted by Conneau et al. (2020b) and Muller et al. (2021). Conneau et al. (2020b) show that the CKA similarity is highly correlated with sentence retrieval scores for four languages. In this paper, our focus is not on finding different ways to calculate similarity between languages, but on how helpful a properly defined alignment score can be in estimating the multilingual capabilities of LLMs across multiple languages.

A.4 ROBUSTNESS OF MEXA

We show that the MEXA alignment score ($\mu(\cdot)$) is very robust, and the odds of this score randomly achieving a high value are very slim. Recall that $\mu(C(L_1, L_2, m, l))$ measures the fraction of diagonal elements in matrix $C(L_1, L_2, m, l)$ that have the maximum value in their respective rows and columns. If this condition is met k times out of n diagonal elements, then $\mu(C(L_1, L_2, m, l))$ is $\frac{k}{n}$. In an $n \times n$ random matrix, the probability of a diagonal element being the maximum in its row and column (a total of $2n - 1$ elements) is $p = \frac{1}{2n-1}$. The probability that at least k out of n independent variables are satisfied, given that the diagonal element is the maximum in its row and column, can be computed using the binomial distribution with Eq. 3.

$$P(X \geq \frac{k}{n}) = 1 - \sum_{i=0}^{k-1} \binom{n}{i} p^i (1-p)^{n-i} \quad (3)$$

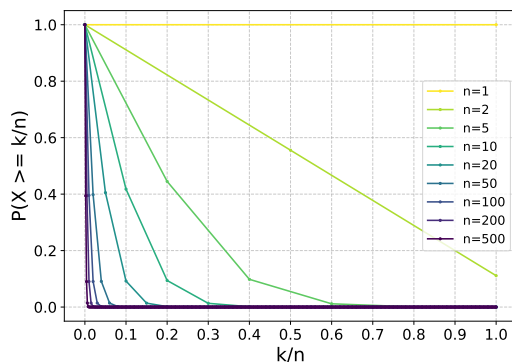


Figure 4: The probability that at least k out of n diagonal elements in an $n \times n$ random matrix are the maximum elements in their respective rows and columns.

In Figure 4, we plot $P(X \geq \frac{k}{n})$. This plot illustrates that, given a sufficient number of parallel sentences (n), the probability of achieving a high score by chance is very low. For example, with $n = 100$, the chance of obtaining MEXA alignment score larger than 0.05 ($k = 5$) from a 100×100 random matrix is $P(X \geq 0.05) = 0.00016$.

A.5 MEXA FOR FLORES-200

We compute MEXA with weighted average embedding and max pooling for the FLORES parallel data for 203 language labels, multiplied by the performance of Belebele for each model in English. We show the results in Table 5, and color the cells based on 0.2 intervals from green (well-covered) to red (not covered): (1.0-0.8), (0.8-0.6), (0.6-0.4), (0.4-0.2), (0.2-0). Note that although FLORES is a high-quality, human-translated dataset, we addressed two major issues before proceeding, as noted by Kargaran et al. (2023). First, the data labeled as Cantonese (Yue Chinese) is not actually Cantonese, so we removed it. Second, the code for Central Atlas Tamazight (tzm), which actually refers to Standard Moroccan Tamazight (zgh), was renamed accordingly. As Belebele is relatively an easy task since the models get good scores in English, and we are using max pooling, this gives a high estimate of the coverage the LLMs have. If the score for a language is not very high, it likely indicates that for more challenging tasks, it will remain low. In Table 5, we can clearly see that Llama 3.1-70B and Gemma 2-9B show a higher level of multilinguality than other models.

	Gemma 2 9B	Gemma 1 7B	Llama 3.1 70B	Llama 3.1 8B	Llama 3 8B	Llama 2 7B	Llama 1 7B	Mistral 7B	OLMo 7B	AVG
eng_Latn	0.92	0.85	0.95	0.88	0.87	0.48	0.42	0.84	0.77	0.77
fra_Latn	0.92	0.84	0.94	0.88	0.87	0.37	0.41	0.84	0.70	0.75
por_Latn	0.92	0.84	0.94	0.88	0.87	0.41	0.41	0.84	0.63	0.75
deu_Latn	0.92	0.84	0.94	0.88	0.87	0.35	0.42	0.84	0.65	0.74
spa_Latn	0.92	0.83	0.95	0.88	0.87	0.37	0.42	0.84	0.56	0.74
ita_Latn	0.92	0.83	0.92	0.88	0.87	0.35	0.42	0.84	0.56	0.73
cat_Latn	0.92	0.82	0.94	0.88	0.87	0.39	0.42	0.84	0.50	0.73
nld_Latn	0.92	0.82	0.95	0.88	0.87	0.34	0.42	0.84	0.52	0.73
rus_Cyrl	0.91	0.82	0.94	0.88	0.87	0.34	0.41	0.83	0.51	0.72
zho_Hans	0.91	0.80	0.94	0.88	0.87	0.32	0.34	0.81	0.62	0.72
glg_Latn	0.92	0.83	0.91	0.88	0.87	0.31	0.41	0.82	0.52	0.72
swe_Latn	0.92	0.83	0.95	0.88	0.87	0.38	0.42	0.84	0.37	0.72
dan_Latn	0.92	0.83	0.94	0.88	0.87	0.31	0.41	0.82	0.44	0.71
ces_Latn	0.92	0.82	0.95	0.88	0.87	0.26	0.42	0.84	0.43	0.71
ron_Latn	0.92	0.82	0.94	0.88	0.87	0.23	0.41	0.83	0.48	0.71
nob_Latn	0.91	0.82	0.95	0.88	0.87	0.34	0.39	0.81	0.39	0.71
zho_Hant	0.91	0.81	0.94	0.88	0.87	0.31	0.32	0.79	0.52	0.71
pol_Latn	0.92	0.81	0.95	0.88	0.87	0.22	0.42	0.84	0.38	0.70

Continued on next page

	Gemma 2	Gemma 1	Llama 3.1	Llama 3.1	Llama 3	Llama 2	Llama 1	Mistral	OLMo	AVG	
	9B	7B	70B	8B	8B	7B	7B	7B	7B		
1188											
1189											
1190											
1191	ast_Latn	0.90	0.80	0.91	0.88	0.86	0.21	0.40	0.77	0.49	0.69
1192	ind_Latn	0.92	0.83	0.93	0.87	0.87	0.22	0.30	0.82	0.42	0.69
1193	oci_Latn	0.89	0.75	0.95	0.88	0.87	0.22	0.39	0.81	0.40	0.68
1194	bos_Latn	0.91	0.81	0.95	0.88	0.87	0.19	0.41	0.84	0.25	0.68
1195	nno_Latn	0.92	0.82	0.92	0.84	0.84	0.26	0.36	0.78	0.38	0.68
1196	ukr_Cyrl	0.92	0.81	0.95	0.88	0.87	0.22	0.42	0.84	0.15	0.67
1197	zsm_Latn	0.92	0.83	0.93	0.88	0.87	0.17	0.25	0.81	0.36	0.67
1198	hrv_Latn	0.91	0.81	0.90	0.86	0.86	0.18	0.41	0.83	0.23	0.67
1199	slv_Latn	0.91	0.79	0.93	0.86	0.86	0.20	0.40	0.84	0.19	0.66
1200	afr_Latn	0.91	0.81	0.93	0.87	0.87	0.20	0.37	0.79	0.21	0.66
1201	slk_Latn	0.91	0.80	0.93	0.86	0.85	0.12	0.38	0.82	0.25	0.66
1202	bul_Cyrl	0.91	0.80	0.90	0.86	0.86	0.12	0.42	0.84	0.14	0.65
1203	jpn_Jpan	0.90	0.80	0.93	0.83	0.82	0.29	0.25	0.76	0.24	0.65
1204	hun_Latn	0.91	0.78	0.92	0.84	0.83	0.13	0.39	0.81	0.18	0.64
1205	vec_Latn	0.87	0.74	0.93	0.84	0.83	0.16	0.35	0.76	0.28	0.64
1206	srp_Cyrl	0.91	0.79	0.90	0.86	0.86	0.10	0.42	0.84	0.06	0.64
1207	tgl_Latn	0.91	0.74	0.94	0.82	0.82	0.16	0.20	0.77	0.36	0.64
1208	fin_Latn	0.91	0.79	0.90	0.85	0.85	0.14	0.21	0.74	0.33	0.64
1209	mkd_Cyrl	0.90	0.77	0.94	0.87	0.86	0.07	0.38	0.80	0.11	0.63
1210	vie_Latn	0.91	0.81	0.95	0.88	0.87	0.22	0.08	0.79	0.16	0.63
1211	epo_Latn	0.87	0.76	0.95	0.86	0.85	0.14	0.26	0.67	0.11	0.61
1212	kor_Hang	0.88	0.74	0.91	0.84	0.83	0.22	0.15	0.71	0.15	0.60
1213	arb_Arab	0.91	0.80	0.94	0.86	0.85	0.05	0.17	0.70	0.10	0.60
1214	ars_Arab	0.91	0.80	0.93	0.86	0.85	0.04	0.17	0.69	0.08	0.59
1215	lim_Latn	0.76	0.65	0.89	0.83	0.83	0.21	0.25	0.59	0.21	0.58
1216	acq_Arab	0.91	0.78	0.92	0.83	0.82	0.04	0.13	0.67	0.09	0.58
1217	acm_Arab	0.90	0.76	0.90	0.86	0.82	0.04	0.14	0.67	0.09	0.57
1218	fur_Latn	0.73	0.60	0.91	0.81	0.77	0.16	0.27	0.59	0.26	0.57
1219	pes_Arab	0.91	0.79	0.88	0.85	0.85	0.05	0.08	0.59	0.07	0.56
1220	arz_Arab	0.88	0.74	0.90	0.84	0.83	0.03	0.10	0.63	0.09	0.56
1221	ajp_Arab	0.88	0.76	0.86	0.85	0.83	0.03	0.12	0.60	0.09	0.56
1222	lit_Latn	0.90	0.76	0.92	0.78	0.80	0.10	0.10	0.56	0.11	0.56
1223	apc_Arab	0.89	0.76	0.86	0.82	0.83	0.03	0.11	0.64	0.09	0.56
1224	ell_Grek	0.90	0.78	0.87	0.87	0.86	0.02	0.09	0.58	0.05	0.56
1225	tur_Latn	0.89	0.78	0.90	0.82	0.81	0.04	0.09	0.61	0.04	0.55
1226	est_Latn	0.90	0.77	0.90	0.82	0.83	0.12	0.09	0.45	0.10	0.55
1227	pap_Latn	0.79	0.60	0.89	0.75	0.73	0.18	0.22	0.56	0.23	0.55
1228	lmo_Latn	0.73	0.56	0.87	0.75	0.74	0.17	0.26	0.60	0.26	0.55
1229	szl_Latn	0.77	0.59	0.87	0.73	0.74	0.11	0.26	0.64	0.21	0.55
1230	prs_Arab	0.90	0.78	0.92	0.84	0.84	0.01	0.06	0.46	0.08	0.54
1231	scn_Latn	0.77	0.59	0.88	0.79	0.77	0.15	0.22	0.57	0.15	0.54
1232	heb_Hebr	0.91	0.81	0.89	0.83	0.83	0.02	0.05	0.47	0.06	0.54
1233	lvs_Latn	0.90	0.75	0.90	0.81	0.79	0.05	0.05	0.55	0.08	0.54
1234	als_Latn	0.87	0.67	0.93	0.79	0.80	0.09	0.08	0.53	0.10	0.54
1235	lij_Latn	0.74	0.58	0.88	0.72	0.70	0.16	0.25	0.53	0.30	0.54
1236	ceb_Latn	0.83	0.59	0.89	0.73	0.72	0.16	0.15	0.49	0.24	0.53
1237	srd_Latn	0.73	0.59	0.86	0.75	0.72	0.16	0.23	0.55	0.21	0.53
1238	hin_Deva	0.90	0.74	0.91	0.80	0.79	0.03	0.05	0.44	0.06	0.53
1239	ltz_Latn	0.79	0.59	0.84	0.75	0.74	0.15	0.18	0.44	0.18	0.52
1240	tha_Thai	0.90	0.76	0.87	0.83	0.83	0.02	0.02	0.32	0.10	0.52
1241	aeb_Arab	0.82	0.67	0.86	0.78	0.75	0.04	0.10	0.55	0.08	0.52
1242	bel_Cyrl	0.88	0.65	0.88	0.79	0.79	0.02	0.09	0.50	0.02	0.51
1243	isl_Latn	0.83	0.62	0.88	0.77	0.78	0.09	0.10	0.48	0.06	0.51
1244	swl_Latn	0.90	0.74	0.86	0.73	0.80	0.11	0.09	0.27	0.08	0.51
1245	mlt_Latn	0.88	0.63	0.87	0.74	0.74	0.12	0.11	0.38	0.12	0.51
1246	war_Latn	0.76	0.55	0.88	0.65	0.61	0.26	0.20	0.35	0.20	0.49
1247	cym_Latn	0.87	0.59	0.88	0.75	0.76	0.11	0.10	0.28	0.08	0.49
1248	fao_Latn	0.71	0.53	0.86	0.71	0.69	0.12	0.13	0.53	0.08	0.48
1249	urd_Arab	0.83	0.66	0.88	0.76	0.73	0.02	0.02	0.31	0.03	0.47
1250	jav_Latn	0.75	0.54	0.84	0.69	0.67	0.16	0.12	0.29	0.16	0.47

Continued on next page

	Gemma 2	Gemma 1	Llama 3.1	Llama 3.1	Llama 3	Llama 2	Llama 1	Mistral	OLMo	AVG
	9B	7B	70B	8B	8B	7B	7B	7B	7B	
1242										
1243										
1244										
1245										
1246										
1247										
1248										
1249										
1250										
1251										
1252										
1253										
1254										
1255										
1256										
1257										
1258										
1259										
1260										
1261										
1262										
1263										
1264										
1265										
1266										
1267										
1268										
1269										
1270										
1271										
1272										
1273										
1274										
1275										
1276										
1277										
1278										
1279										
1280										
1281										
1282										
1283										
1284										
1285										
1286										
1287										
1288										
1289										
1290										
1291										
1292										
1293										
1294										

Continued on next page

	Gemma 2	Gemma 1	Llama 3.1	Llama 3.1	Llama 3	Llama 2	Llama 1	Mistral	OLMo	AVG	
	9B	7B	70B	8B	8B	7B	7B	7B	7B		
1296											
1297											
1298											
1299	sin_Sinh	0.56	0.27	0.49	0.26	0.18	0.01	0.01	0.03	0.02	0.20
1300	sno_Latn	0.28	0.09	0.66	0.20	0.20	0.10	0.08	0.13	0.06	0.20
1301	nya_Latn	0.36	0.13	0.41	0.19	0.19	0.13	0.10	0.17	0.06	0.19
1302	twi_Latn	0.23	0.08	0.46	0.22	0.22	0.14	0.13	0.19	0.07	0.19
1303	sna_Latn	0.41	0.17	0.40	0.19	0.20	0.10	0.08	0.12	0.07	0.19
1304	uig_Arab	0.21	0.09	0.71	0.29	0.29	0.00	0.01	0.03	0.02	0.18
1305	bug_Latn	0.14	0.12	0.35	0.22	0.22	0.14	0.11	0.20	0.12	0.18
1306	luo_Latn	0.07	0.07	0.40	0.25	0.24	0.15	0.12	0.21	0.09	0.18
1307	tsn_Latn	0.24	0.10	0.42	0.18	0.18	0.12	0.11	0.20	0.06	0.18
1308	arb_Latn	0.29	0.07	0.46	0.24	0.20	0.05	0.05	0.17	0.08	0.18
1309	khm_Khmr	0.34	0.15	0.59	0.15	0.16	0.01	0.02	0.09	0.06	0.17
1310	lua_Latn	0.09	0.08	0.33	0.20	0.21	0.14	0.13	0.24	0.12	0.17
1311	lug_Latn	0.17	0.07	0.41	0.18	0.19	0.14	0.09	0.19	0.06	0.17
1312	grn_Latn	0.17	0.09	0.44	0.16	0.17	0.12	0.09	0.13	0.10	0.16
1313	ssw_Latn	0.27	0.10	0.37	0.17	0.17	0.11	0.08	0.14	0.05	0.16
1314	lin_Latn	0.11	0.08	0.43	0.16	0.18	0.12	0.11	0.16	0.08	0.16
1315	ory_Orya	0.28	0.03	0.66	0.18	0.19	0.01	0.01	0.03	0.03	0.16
1316	fij_Latn	0.13	0.06	0.38	0.18	0.16	0.14	0.11	0.15	0.08	0.15
1317	fuv_Latn	0.07	0.08	0.30	0.20	0.20	0.13	0.10	0.18	0.10	0.15
1318	kas_Arab	0.16	0.10	0.50	0.20	0.21	0.02	0.02	0.10	0.05	0.15
1319	quy_Latn	0.10	0.06	0.42	0.21	0.22	0.10	0.06	0.13	0.05	0.15
1320	aka_Latn	0.17	0.06	0.37	0.14	0.17	0.11	0.08	0.12	0.10	0.15
1321	mya_Mymr	0.36	0.13	0.46	0.14	0.16	0.00	0.00	0.02	0.02	0.15
1322	run_Latn	0.25	0.07	0.37	0.16	0.17	0.08	0.06	0.10	0.04	0.14
1323	bem_Latn	0.14	0.08	0.29	0.16	0.16	0.13	0.11	0.15	0.06	0.14
1324	kas_Deva	0.14	0.09	0.37	0.20	0.21	0.02	0.03	0.11	0.08	0.14
1325	wol_Latn	0.09	0.07	0.30	0.18	0.16	0.12	0.10	0.17	0.07	0.14
1326	kam_Latn	0.09	0.08	0.26	0.18	0.16	0.13	0.10	0.15	0.08	0.14
1327	tso_Latn	0.14	0.06	0.35	0.14	0.13	0.11	0.08	0.13	0.06	0.13
1328	kon_Latn	0.07	0.08	0.27	0.15	0.17	0.09	0.07	0.13	0.09	0.13
1329	tum_Latn	0.15	0.07	0.32	0.11	0.13	0.09	0.08	0.11	0.05	0.13
1330	kik_Latn	0.07	0.04	0.32	0.12	0.13	0.10	0.09	0.13	0.12	0.12
1331	taq_Latn	0.06	0.06	0.28	0.14	0.12	0.11	0.08	0.14	0.05	0.12
1332	mos_Latn	0.04	0.04	0.25	0.16	0.14	0.11	0.09	0.15	0.07	0.12
1333	yor_Latn	0.13	0.04	0.30	0.14	0.14	0.08	0.06	0.10	0.04	0.11
1334	amh_Ethi	0.48	0.16	0.24	0.04	0.03	0.01	0.02	0.03	0.02	0.11
1335	sag_Latn	0.05	0.07	0.22	0.17	0.17	0.07	0.07	0.09	0.06	0.11
1336	cjk_Latn	0.06	0.06	0.21	0.13	0.12	0.10	0.08	0.13	0.07	0.11
1337	umb_Latn	0.05	0.05	0.20	0.15	0.14	0.10	0.08	0.11	0.05	0.10
1338	dyu_Latn	0.04	0.04	0.22	0.13	0.12	0.06	0.07	0.10	0.08	0.10
1339	kac_Latn	0.02	0.03	0.22	0.12	0.14	0.06	0.06	0.12	0.08	0.09
1340	kmb_Latn	0.05	0.06	0.20	0.11	0.10	0.10	0.07	0.09	0.05	0.09
1341	bam_Latn	0.05	0.05	0.18	0.11	0.09	0.08	0.08	0.12	0.04	0.09
1342	ayr_Latn	0.04	0.04	0.20	0.11	0.10	0.06	0.05	0.10	0.06	0.08
1343	lao_Laoo	0.17	0.04	0.22	0.07	0.09	0.02	0.02	0.04	0.09	0.08
1344	dik_Latn	0.05	0.06	0.18	0.06	0.07	0.08	0.07	0.10	0.05	0.08
1345	ewe_Latn	0.04	0.03	0.18	0.08	0.08	0.09	0.07	0.08	0.04	0.08
1346	knc_Latn	0.05	0.06	0.15	0.08	0.08	0.07	0.06	0.08	0.05	0.08
1347	kab_Latn	0.04	0.02	0.17	0.09	0.08	0.06	0.06	0.11	0.04	0.07
1348	sat_Olck	0.19	0.02	0.32	0.05	0.05	0.00	0.00	0.01	0.01	0.07
1349	gaz_Latn	0.05	0.03	0.20	0.06	0.06	0.05	0.04	0.08	0.03	0.07
1350	bod_Tibt	0.07	0.01	0.22	0.08	0.08	0.01	0.01	0.02	0.02	0.06
1351	fon_Latn	0.03	0.02	0.14	0.06	0.06	0.03	0.04	0.05	0.07	0.06
1352	shn_Mymr	0.02	0.01	0.21	0.06	0.06	0.01	0.02	0.03	0.07	0.05
1353	kbp_Latn	0.03	0.02	0.14	0.05	0.04	0.03	0.02	0.08	0.05	0.05
1354	mni_Beng	0.03	0.02	0.12	0.05	0.06	0.01	0.02	0.08	0.02	0.05
1355	ace_Arab	0.03	0.02	0.15	0.07	0.07	0.00	0.00	0.03	0.01	0.04
1356	knc_Arab	0.01	0.01	0.13	0.04	0.04	0.02	0.02	0.05	0.05	0.04
1357	bjn_Arab	0.03	0.02	0.11	0.05	0.08	0.01	0.01	0.05	0.01	0.04
1358	nus_Latn	0.02	0.02	0.07	0.04	0.03	0.03	0.02	0.04	0.05	0.03

Continued on next page

	Gemma 2 9B	Gemma 1 7B	Llama 3.1 70B	Llama 3.1 8B	Llama 3 8B	Llama 2 7B	Llama 1 7B	Mistral 7B	OLMo 7B	AVG
min_Arab	0.02	0.01	0.13	0.05	0.04	0.01	0.00	0.03	0.01	0.03
tir_Ethi	0.10	0.02	0.05	0.02	0.02	0.01	0.01	0.02	0.02	0.03
dzo_Tibt	0.01	0.00	0.08	0.03	0.03	0.00	0.00	0.01	0.01	0.02
taq_Tfng	0.00	0.00	0.04	0.01	0.01	0.00	0.00	0.01	0.03	0.01
zgh_Tfng	0.00	0.00	0.02	0.01	0.01	0.00	0.00	0.01	0.01	0.01

Table 5: Adjusted performance of MEXA using max pooling with the English performance of models on the Belebele benchmark.

1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403