The BELEBELE Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants

Anonymous ACL submission

Abstract

We present BELEBELE, a multiple-choice machine reading comprehension (MRC) dataset spanning 122 language variants. Significantly expanding the language coverage of natural language understanding (NLU) benchmarks, this dataset enables the evaluation of text models in high-, medium-, and low-resource languages. Each question is based on a short passage from the FLORES-200 dataset and has four multiple-choice answers. The questions were carefully curated to discriminate between models with different levels of general language comprehension. The English dataset on its own proves difficult enough to challenge state-of-the-art language models. Being fully parallel, this dataset enables direct comparison of model performance across all languages. We use this dataset to evaluate the capabilities of multilingual masked language models (MLMs) and large language models (LLMs). We present extensive results and findings, notably that despite significant cross-lingual transfer in English-centric LLMs, much smaller MLMs pretrained on balanced multilingual data still understand far more languages. Overall, BELEBELE opens up new avenues for evaluating and analyzing the multilingual capabilities of NLP systems.

1 Introduction

The absence of high-quality, parallel evaluation benchmarks is a major obstacle in assessing the text comprehension capabilities of multilingual models. NLP datasets with high language coverage do exist, such as FLORES-200 (NLLB et al., 2022), but they primarily focus on machine translation. Popular multilingual evaluation benchmarks, such as multilingual question answering (Lewis et al., 2020; Clark et al., 2020), natural language inference (NLI) (Conneau et al., 2018), and summarization (Ladhak et al., 2020; Hasan et al., 2021), altogether only cover around 30 languages. And while understanding and generative text services are used across the globe in 100+ languages, the lack of labeled data provides a major obstacle to building functional systems in most languages.

Simultaneously, large language models (LLMs) have become increasingly popular. Certain LLMs, like BLOOM (Scao et al., 2022), are trained on multilingual data and tout their innate multilingual capabilities. Others like GPT-3 (Brown et al., 2020) and LLAMA (Touvron et al., 2023a) have demonstrated multilingual competence despite their training data being predominantly English-centric. Even so, LLMs benefit from pretraining data that is linguistically diverse, intentionally or not, as well as from cross-lingual transfer (Zoph et al., 2016; Artetxe et al., 2020; Muller et al., 2021b). But how multilingual are these models really? Beyond LLMs, significant scientific progress needs to be made before NLP systems can be built effectively and efficiently in low-resource languages. Many modeling techniques are being presented as language-*agnostic* but have only truly been evaluated in a small number of languages (Bender, 2011), risking not being applicable to diverse typologically phenomena (Bender, 2009). We believe that large-scale, parallel, and discriminative datasets are crucial for studying the multilingual capabilities of such models and understanding how the technological disparity between high- and low-resource languages is evolving.

In this paper, we present a fundamental natural language understanding benchmark to evaluate language models across 122 language variants from around the world, called BELEBELE[1]. The dataset contains 900 unique multiple-choice reading comprehension passages and questions. The questions have been carefully crafted to discriminate between models with varying competence in language comprehension. While the questions do not necessarily require higher levels of knowledge or reasoning, they favor generalizable NLU models and deliberately punish biased models. The English questions

[1] Bambara word meaning "big, large, fat, great".

Language	Passage	Multiple-choice questions			
English eng_Latn	Insects were the first animals to take to the air. Their ability to fly helped them evade enemies more easily and find food and mates more efficiently. Most insects have the advantage of being able to fold their wings back along the body. This gives them a wider range of small places to hide from predators. Today, the only insects that cannot fold back their wings are dragon flies and mayflies.	Q1. An insect's ability to fold their wings back increases which of the following?			
		1. Food supply	2. Hiding spaces	3. Finding mates	4. Flight speed
		Q2. Which of the following is not mentioned in the passage as something affected by an insect's flying ability?			
		1. Reproduction	2. Life span	3. Food gathering	4. Efficiency
Tigrinya tir_Ethi	ሓሽራ ናብ ኣየር ዝወሰዱ ኣፈናዊ እንስሳ ነይሩ። ናይ ምንፍር ክእለቶም ጸላእቲ ብቐሊል ከጥቅዑን ከምኡውን ምግቢን መሓዛ ብብኹዓት ክረኽቡ ሓጊዝዎም። መብዛሕቲኦም ሓሳኹ መንገብገቦም ናብ ኣነላቶም ናይ ምዕጻፍ ክእለት/ብልጫ ኣለዎም። እዚ ካብ ሃደንቲ ክሕብኡሎም ዝኽእሉሎም ብዙሓት እናኸሸተይ ቦታታት ዝህቦም እዩ። ኣብ'ዚ ሕጂ እዋን፣ እቶም መንገብገቦም ክዓጽፉ ዘይክእሉ እንኮ ሓሳኹ ደራጎንፍላይስን ሜይፍላይስን እዮም።	Q1. ክእለት ናይ ሓደ ሓሳኻ መንገብገቦም ንድሕሪት ክዓጽፉ ምኽኣል ካብዞም ዝስዕቡ ነየኖት ይጠሰኽ?			
		1. ቐረብ መግቢ	2. መሕቢኢ ቦታታት	3. ምርካብ መጻምዲ	4. ናይ ምህዳም ፍጥነት
		Q2. ካብዞም ዝስዕቡ ኣብቲ መግለጺ ከም ሓደ ነገር ብክእለት ምንፋር ናይ ሓሳኻ ዝጽለዉ ተባሂሉ ዘይተጠቕሰ ኣየናይ እዩ?			
		1. ምፍራይ	2. ንዉሓት ሂወት	3. ምእካብ መግቢ	4. ብኹዓት
Khmer khm_Khmr	សត្វល្អិត គឺជាសត្វដំបូងគេដែលហោះហើរនៅក្នុងខ្យល់។ សមត្ថភាពរបស់ពួកគេក្នុងការហោះហើរ បានជួយពួកវាឲ្យល្អានពានសត្រូវជាច្រើនបានដោយងាយស្រួល និងធ្វើឲ្យការស្វែងរកចំណីនិងការបន្តពូជកាន់មានភាពងាយស្រួល។ គុណសម្បត្តិនៃសត្វល្អិតភាគច្រើនគឺ ពួកវាអាចបត់ស្លាបត្រឲ្យមកស្របតាមដងខ្លួនរបស់វាវិញ។ ការនេះផ្តល់ឱ្យពួកវានូវកន្លែងតូចៗចម្រុះជាច្រើនដើម្បីលាក់ខ្លួនពីពួកមំសាសី។ សត្វថ្ងៃនេះ សត្វល្អិតតែមួយគត់ដែលមិនអាចបត់ស្លាបរបស់ខ្លួនបាន គឺសត្វកន្ទុំរុយចម្មតា និងកន្ទុំរុយស្លាបរមេអំបៅ។	Q1. សមត្ថភាពរបស់សត្វល្អិតក្នុងការបត់ស្លាបត្រឡប់មកស្របតាមដងខ្លួនបង្កើនមួយណាដូចខាងក្រោម?			
		1. ការផ្គត់ផ្គង់ស្បៀងអាហារ	2. កន្លែងលាក់ខ្លួន	3. ការស្វែងរកការបន្តពូជ	4. ល្បឿនហោះហើរ
		Q2. តើអតីតរដ្ឋាភិបាលអូស្ត្រាលីបានគិតថា សម្ភាប៍ឆនទីក្រុងក្យូតនឹងបណ្តាលឱ្យមានការខូចខាតអ្វីខ្លះ?			
		1. ការបន្តពូជ	2. អាយុកាល	3. ការស្វែងរកអាហារ	4. ប្រសិទ្ធភាពការហោះហើរ
Portuguese por_Latn	Insetos foram os primeiros animais a irem para o ar. Sua habilidade de voar os ajudou a fugir mais facilmente de inimigos e a encontrar alimento e parceiros de maneira mais eficiente. A maioria dos insetos têm a vantagem de poder dobrar suas asas ao longo do corpo. Isso lhes dá uma gama maior de pequenos locais para se esconderem de predadores. Hoje, os únicos insetos que não podem dobrar suas asas para trás são libélulas e efêmeras.	Q1. A capacidade de um inseto de dobrar suas asas para trás aumenta qual das seguintes opções?			
		1. Suprimento de comida	2. Esconderijos	3. Encontrar parceiros/as	4. Velocidade de voo
		Q2. Qual das seguintes opções não é mencionada no trecho como algo afetado pela habilidade de voar de um inseto?			
		1. Reprodução	2. Vida útil	3. Obtenção de alimentos	4. Eficiência

Figure 1: A sample passage from the dataset in 4 different languages, displayed alongside its two questions.

on their own present a significant challenge to numerous models, while humans are capable of answering the questions with near-perfect accuracy.

The first of its scale, BELEBELE is parallel across all languages, facilitating a direct comparison of model performance across all languages. The dataset covers typologically diverse languages across 29 scripts and 27 language families. Seven languages are included in two separate scripts, resulting in one of the first NLP benchmarks for the romanized variants of Hindi, Urdu, Bengali, Nepali, and Sinhala. We further detail our data collection process and the resulting corpus in Section 3.

The dataset enables evaluation of mono- and multi-lingual models, but the parallel nature also enables a number of cross-lingual evaluation settings. We evaluate several masked language models (MLMs) after fine-tuning on an English training set as well as with the assistance of machine translation (Translate-Train-All). For LLMs, we evaluate several models using In-Context Learning and also instruction-tuned models via Zero-Shot. We discuss our results in Section 5.

2 Background

2.1 Cross-Lingual Evaluation Benchmarks

There are several datasets for NLU that are parallel across numerous languages and enable monolingual, multilingual, or cross-lingual evaluation. These include XNLI (Conneau et al., 2018), XQuAD (Artetxe et al., 2020), and MLQA (Lewis et al., 2020). MINTAKA (Sen et al., 2022) is designed with LLMs in mind, presenting a more difficult QA task in 9 languages. Beyond QA, XL-SUM (Hasan et al., 2021) is an analogous dataset in the domain of abstractive summarization. However, all these datasets together cover under 30 languages, most of which are high- or medium- resource. MASSIVE (FitzGerald et al., 2023) is a large NLU dataset covering 51 languages, but in the domain of spoken conversational agents. NER (Pan et al., 2017) has extensive language coverage and TYDIQA (Clark et al., 2020) is a popular multilingual benchmark but neither are parallel.

Our work undertakes the challenge of expanding existing cross-lingual evaluations to 122 languages, many of which currently lack any NLU benchmark at all.

2.2 Non-English Machine Reading Comprehension

While the question-answering portion varies, machine reading comprehension (MRC) tasks are defined by the closed-book passage provided to answer each question. Of course, a big majority of MRC datasets are in English, such as TRIVIAQA (Joshi et al., 2017) and the BABI tasks (Weston et al., 2016).

However, the need for MRC datasets for other languages has led to a proliferation of monolingual closed-book MRC datasets in recent years (Mozannar et al., 2019; Hardalov et al., 2019; d'Hoffschmidt et al., 2020; Möller et al., 2021;

BELEBELE Statistics					
Languages		Passage statistics		Question statistics	
Total Number	122	Distinct Passages	488	Distinct Questions	900
Distinct Languages (ignoring script)	115	Questions per passage	1-2	Multiple-choice answers (num correct) per question	4 (1)
Language Families	27	Avg. words per passage (std)	79.1 (26.2)	Avg. words per question (std)	12.9 (4.0)
Scripts	29	Avg. sentences per passage (std)	4.1 (1.4)	Avg. words per answer (std)	4.2 (2.9)

Table 1: Language and Text Information for BELEBELE. Average length statistics are computed on the English split.

Anuranjana et al., 2019; Gupta et al., 2018; Croce et al., 2018; Efimov et al., 2020; Shavrina et al., 2020; Sun et al., 2021). Most were created using translation and so are parallel with an English dataset, often SQUAD (Rajpurkar et al., 2016). However, BELEBELE aims to cover these languages and more in one consistent dataset.

2.3 Multiple Choice QA

Compared to extractive QA, multiple-choice is a less common form of MRC datasets. Some, like RACE (Lai et al., 2017), are made from exam questions for English learners, while others were built specifically for NLU systems, like MCTest (Richardson et al., 2013) and MultiRC (Khashabi et al., 2018). While most are intended to be closed-book, SCIQ (Welbl et al., 2017) and OPENBOOKQA (Mihaylov et al., 2018) require open information retrieval. Others, like COPA (Roemmele et al., 2011), SWAG (Zellers et al., 2018), and RECLOR (Yu et al., 2020), require higher-level commonsense reasoning to answer. For multilingual systems, EXAMS (Hardalov et al., 2020) is a parallel multiple-choice QA dataset covering 28 languages. However, no passages are provided and answering questions requires cross-lingual knowledge transfer and reasoning.

2.4 FLORES-200

The passages in the BELEBELE corpus are directly sourced from the FLORES-200 Machine Translation Benchmark (Goyal et al., 2022; NLLB et al., 2022). The parallel dataset was constructed by sourcing English passages from Wikinews, Wikijunior, and WikiVoyage. The translations were performed by native speakers with high English fluency and translation experience. Translators were instructed to maintain informative and standardized content while handling named entities, abbreviations, idioms, and pronouns appropriately.

3 The BELEBELE Dataset

We opted to create multiple-choice questions and answers in English and then translate, as opposed to creating resources natively in each language. Many of the advantages to this approach outlined in Conneau et al. (2018) remain. Most importantly, this leads to significantly more similar sets of samples across languages, enabling direct score comparison. The process for creating the dataset is summarized in Fig. 2.

3.1 Creation of Multiple Choice Questions & Answers

To create the BELEBELE dataset, we first construct a question-answering dataset in English.

Amongst machine reading comprehension tasks, we select multiple-choice questions (MCQs) because it would lead to the fairest evaluation across languages. Related tasks, such as span extraction, are more sensitive to morphological differences, making scaling to many languages difficult (Lewis et al., 2020). In addition, MCQs enable us to better center the questions on information explicitly stated in the passage, as yes/no or entailment (NLI) questions can be easier to answer with external knowledge held in pretrained models. In order for the questions to discriminate solely between different levels of language understanding, we intentionally create questions that do not require higher levels of information processing, such as multi-hop or commonsense reasoning.

Constructing high quality MCQs depends most importantly on creating strong negatives that are neither obviously wrong nor possibly correct (Agarwal and Mannem, 2011; Richardson et al., 2013). We do not want the dataset to be easy enough for biased models (e.g. models that use shortcuts or pattern-match) (Boyd-Graber and Börschinger, 2020). In setting up this annotation, we consider the protocols proposed in Bowman et al. (2020) and the warnings from Malaviya et al. (2022). We implement an iterative procedure with the Language Service Provider (LSP) for this involved data collection task, similar to that from Nangia et al. (2021). We engaged in 5 total iterations, providing and receiving feedback in each. Annotators were instructed on the similarities and differences on how ML models approach QA datasets versus humans, which we felt substantially improved the quality of the data.

Our final guidelines include both important points such as having the correct response being

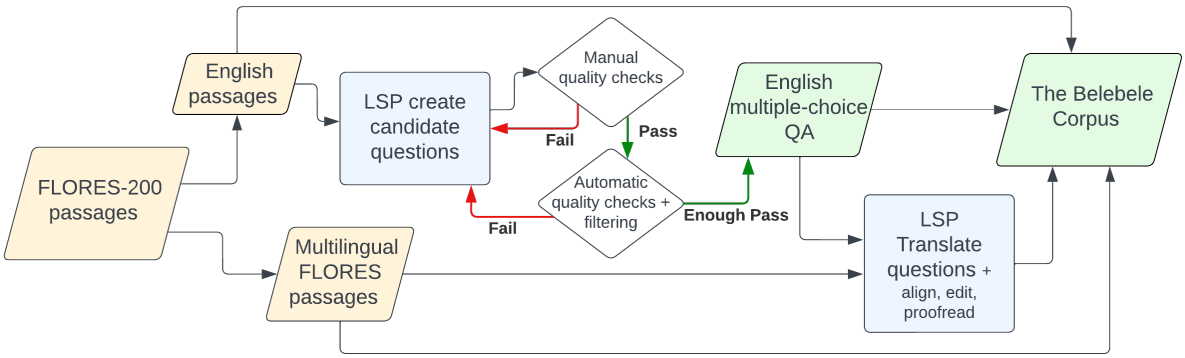

Figure 2: Illustration of the dataset creation process with Language Service Provider (LSP) starting from FLORES

unambiguous, as well as particularized rules such as *no double negatives* (Mihaylov et al., 2018). For each rule we provided annotators with a good and bad example to illustrate. An abridged version of our guidelines can be found in the Appendix A.5.1.

3.2 Quality Assurance

At each iteration, we evaluate whether or not returned samples satisfy the minimum quality bar through a mix of manual inspection and automatic inspection. At every step, we manually verified a sample of questions to understand how well the annotators were on the same page with us about the guidelines. While time consuming, manual verification was the most assured way to provide tangible feedback to the annotators, notably on the difficulty of the questions created. As we progressively aligned with annotators, we were required to look over more samples to provide feedback.

To complement the manual inspection of a subset of questions, we use programmatic methods to evaluate all questions from a statistical perspective. Based on the findings in Malaviya et al. (2022), we create low-level features to identify overly easy questions or low-effort strategies employed by annotators. For example, we evaluate the lexical overlap between different combinations of the texts associated with a question to evaluate whether the question is answerable by a biased model. This allows us to see if the question can be answered without the passage, without the question, or with only one sentence in the passage. We also identified patterns associated with heuristic solvability, such as the wrong answers less frequently being extracted from the passage.

These low-level features allow us to (1) determine whether an annotation iteration was up to par, (2) filter out questions that failed these heuristic checks (for the final iteration, about 20% were filtered out), and (3) compare to other MCQ datasets. We run statistical t-tests to ensure the distribution of these features for correct answers is no different than for wrong answers. In comparison to MCTEST which largely fails this t-test (p-value < 0.01), our final collection has p-value 0.81. We also train a logistic regression model to answer using only bag-of-word representations and find that the best the naïve model could achieve was an accuracy of 0.28 on our 900 questions. This is just better than random (0.25) and much lower than what was achieved on MCTEST, 0.44.

3.3 Translating the Corpus

BELEBELE was created end-to-end without the use of machine translation technology, relying solely on experts fluent in English and the target language.

For all languages included in the corpus, the context passages were taken directly from the FLORES-200 dataset, with the exception of Hindi, Bengali, Urdu, Nepali, and Sinhala in the Latin script. While the romanized variant of these 5 Indo-Aryan languages is very prevalent on the modern Internet, their romanization is not included in FLORES-200. We thus had annotators transliterate from the native to Latin script with the support of IndicXlit (Madhani et al., 2023). As a result, much like Modern Standard Arabic, these languages are present in two forms in the corpus.

In order for the questions and answers to properly pair the translated FLORES passages, the latter was provided for the annotators. We specifically instructed annotators to align potentially ambiguous translations with the original passages. While Clark et al. (2020) warns that this forced alignment could increase 'translationese', it is necessary to ensure equivalent question difficulty across languages. The modifications to the translation guidelines can be found in Appendix A.5.2. All translations were proofread and edited by an additional annotator.

Model	Size/Variant	Vocab size	AVG	% ≥ 50	% ≥ 70	eng_Latn	non-Eng AVG
5-Shot In-Context Learning (examples in English)							
Llama 1	7B	32K	27.7	0.0%	0.0%	37.3	27.6
Llama 1	13B	32K	30.4	0.8%	0.0%	53.3	30.2
Llama 1	30B	32K	36.2	18.0%	0.8%	73.1	35.9
Llama 1	70B	32K	40.9	25.4%	12.3%	82.5	40.5
Llama 2 base	70B	32K	**48.0**	**38.5%**	**26.2%**	**90.9**	**47.7**
Falcon	40B	65K	37.3	16.4%	1.6%	77.2	36.9
Zero-Shot for Instructed Models (English instructions)							
Llama-2-chat	7B	32K	34.4	4.1%	0.0%	58.6	34.1
Llama-2-chat	70B	32K	41.5	27.0%	2.5%	78.8	41.2
GPT3.5-turbo	unk	100K	**51.1**	**44.2%**	**29.2%**	**87.7**	**50.7**
Full Finetuning in English							
XLM-R	large (550M)	250K	54.0	64.8%	15.6%	76.2	53.8
XLM-V	large (1.2B)	902K	55.6	**69.7%**	21.2%	76.2	54.9
InfoXLM	large (550M)	250K	**56.2**	67.2%	**28.7%**	**79.3**	**56.0**
Translate-Train-All							
XLM-R	large (550M)	250K	58.9	69.7%	36.1%	78.7	58.8
XLM-V	large (1.2B)	902K	**60.2**	**76.2%**	32.8%	77.8	**60.1**
InfoXLM	large (550M)	250K	60.0	70.5%	**36.9%**	**81.2**	59.8

Table 2: Summary of results on Belebele across models and evaluation settings. % ≥ 50/70 refers to the proportion of languages for which a given model performs above 50/70%. We additionally report Llama-2-chat zero-shot results leveraging translation in Table 3.

3.4 English Training Data

Belebele is intended to be used as a test set, and not for training. Therefore, for models that require additional task finetuning, we instead assemble a training set consisting of samples from English multiple-choice QA datasets (See Appendix A.2).

3.5 The Belebele Dataset in Summary

Belebele contains 900 questions, each with 4 multiple-choice answers and one correct answer. Most passages have two associated questions, but some have only one. In total, there are 488 distinct passages, none belonging to the hidden FLoRes test set. Parallel across 122 languages, the corpus contains a total of 109,800 rows. We display a sample passage in four languages in Fig. 1.

Because of the careful annotation procedure and quality checks, the MCQs discriminate text comprehension competence. It often includes paraphrasing and strong negatives in order to elude simple pattern-matching models. Questions often additionally require understanding multiple sentences. However, answering does not require presumptions or external knowledge as is required in more difficult reasoning datasets. For example, Q1 in Fig. 1 is unambiguous. *Food*, *mates*, and *flying* are all mentioned in the passage, but a careful read reveals the wings folding back is only associated with *hiding spaces*. To confidently rule out the other candidate answers, it is required to understand three sentences. In general, we find all questions to be answerable by humans fluent in the target language, but not without focused reading (see Section 5.1).

As can be seen in Fig. 1, the passages, questions, and answers are aligned in semantic meaning and formality. Belebele therefore poses an equivalent challenge in all languages. It also enables models with cross-lingual alignment in the semantic representation space to answer questions when passage, question, and answers are swapped to different languages. Since FLoRes includes passages in 83 additional languages, we can even evaluate reading comprehension in these languages by asking the questions in English.

4 Experiments

Thanks to Belebele, we are able to evaluate numerous models and establish baseline performances across 122 language variants. We compare performance between popular multilingual MLMs and LLMs in several settings. For all, accuracy is the central metric. With 4 candidate answers for each question, the expected accuracy for sequence classification models that guess randomly is 0.25.[2]

4.1 Evaluated Models

Masked Language Models (MLMs) We evaluate three different models, XLM-V (Liang et al.,

[2]Note: For sequence-to-sequence models (e.g. instructed models) that are evaluated in exact-match scenarios, this lower-bound does not hold.

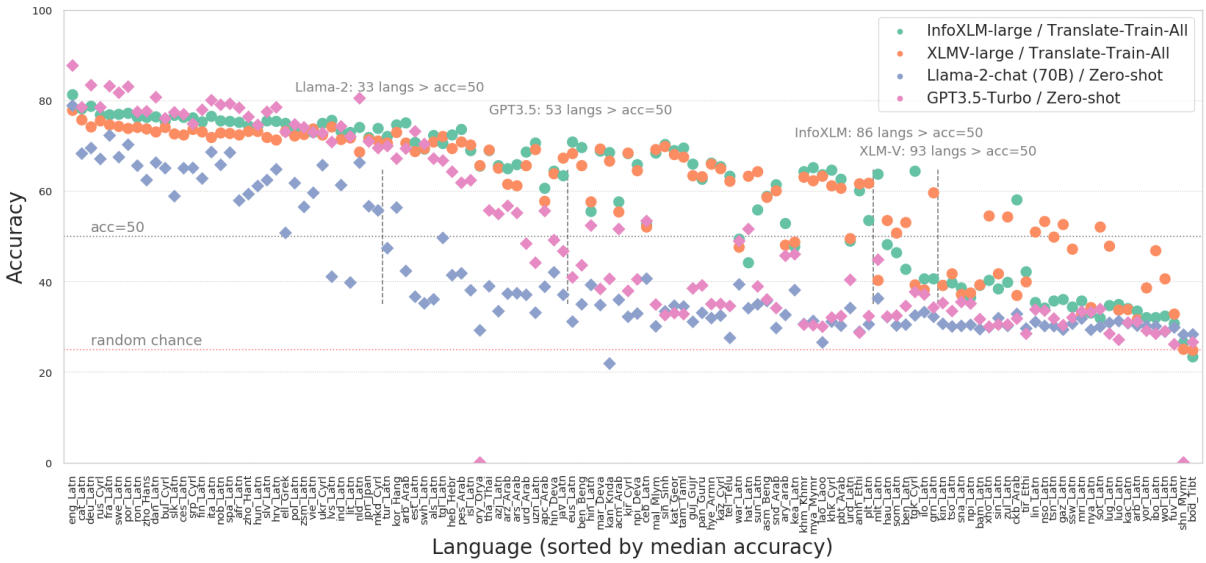

Figure 3: BELEBELE Results in 122 languages. We compare four models in two settings and see the difference between intentionally multilingual models and models with English-centric data. GPT3.5-TURBO performs the best on the top 20 languages, but after 40-50, its performance falls far behind INFOXLM and XLM-V. Similarly, INFOXLM outperforms XLM-V in the first 40 languages, but XLM-V proves more capable on the long tail of languages. Note that the language order can change the plot considerably, here we choose median accuracy.

2023), INFOXLM (Chi et al., 2021), and XLM-R (Conneau et al., 2020a). All the evaluated MLMs have been pretrained on intentionally multilingual corpi inclusive of about 100 languages. The pretraining data in high-resource languages is typically down-sampled while low-resource languages are up-sampled in order to favor multilingual performance (Conneau et al., 2020a). In addition, all their subword tokenizers (Kudo and Richardson, 2018) are trained on multilingual corpora making them better suited for multilingual text.

Large Language Models We evaluate GPT3.5-TURBO, FALCON, and LLAMA (1 and 2). GPT3.5-TURBO is a model optimized for chat based on GPT-3 (Brown et al., 2020) available through OpenAI APIs[3]. Limited details have been disclosed about the pretraining and fine-tuning data.[4] LLAMA 1 (Touvron et al., 2023a) is a collection of decoder-only transformers models trained on 1T (for 7B, 13B) or 1.4T (for 30B, 65B) tokens of web-crawled data, while LLAMA 2 (Touvron et al., 2023b) pretrained on about 2T. We evaluate all four pretrained checkpoints for LLAMA 1. We evaluate both LLAMA 2 70B's pretrained version and its chat version instruction-fine-tuned for safe dialog purposes (a.k.a. LLAMA-2-CHAT). We also evaluate FALCON 40B, which was pretrained on 1T extensively filtered web-crawled samples (Penedo et al., 2023).

While LLAMA 1 was reportedly trained in 20 languages with Latin and Cyrillic scripts, non-English text accounts for less than 4.5% of the pretraining corpus (Touvron et al., 2023a). LLAMA 2 pretraining data is made of 89.7% of English data, 8.4% unidentified, and a tiny 1.9% belonging to 26 other languages[5] (Touvron et al., 2023b). Both series use the same BPE-based tokenizers (Kudo and Richardson, 2018). Splitting unicode characters into bytes also helps LLAMA avoid out-of-vocabulary errors.

4.2 Evaluation Settings

More specifics are provided in Appendix A.4.

Full Model Fine-tuning For evaluating MLMs, we add a multiple-choice classification head and fine-tune the entire model. We finetune in two settings, (1) in English and evaluate cross-lingual transfer and (2) on machine-translated samples of the training set to all the target languages and evaluate each language (*Translate-Train-All*).

Five-Shot In-Context Learning We evaluate the pretrained LLAMA 1 and 2 as well as FALCON 40B in the five-shots setting. Examples are sampled from the English training set and prompted to the model. For prediction, we pick the answer with highest probability and report the average cumulative score over 3 runs.

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

[4] Our experiments rely on the unverified assumption that GPT3.5-TURBO was not trained on FLORES. See the Ethics Statement

[5] See Table 10 in Touvron et al. (2023b) for a full list of the identified languages

Zero-Shot Evaluation We evaluate GPT3.5 and LLAMA-2-CHAT (70B) in Zero-Shot by describing the task with natural language instructions (in English). We present the passage, question, and four possible answers, and instruct the model to provide the letter of the answer. We post-process answers and accept multiple formats.[6]

In addition, we prompt LLAMA-2-CHAT with instructions that are machine translated to the target language from English. Conversely, we evaluate machine-translating the passages, questions, and answers back to English and prompting them to the model (*Translate-Test*). This setting allows us to compare in-language comprehension to the popular approach of cascading with machine translation.

5 Results

We provide summarized results in Table 2 and detailed results in Appendix A.6.

5.1 How difficult is BELEBELE?

As discussed in Section 3, the questions in BELEBELE are intentionally difficult. While the primary challenge of this dataset is its multilinguality, we see that empirically, the English questions are able to shed light on the varying NLU capabilites of models. With full finetuning, we achieved a maximum accuracy of 71.7 in English with ROBERTA-base model, significantly less than the 90.9 achieved by LLAMA 2 70B in five-shot. Between LLAMA 1 variants, we see a wide range of results, with the 7B model only achieving 37.3.

To establish human performance, 4 authors each randomly sampled around 30 English MCQs and answered in a blind test, achieving mean 97.6 accuracy[7]. This is much higher than any of the models evaluated, implying the task presents a particular challenge for models. For comparison, Nangia and Bowman (2019) conservatively estimate human performance to be 92.8 on the English split of XNLI (i.e. MNLI (Williams et al., 2018)).

When comparing model performance with XNLI, we find very high correlation. In the Translate-Train-All setting, XLM-V, INFOXLM, and XLM-R all perform about 10 accuracy points lower on BELEBELE than on XNLI Translate-Train[8] reported in Liang et al. (2023) and Chi et al. (2021). Still, across all 15 languages and three models, we find a score correlation of $r = 0.85$.

[6]We provide code to replicate at our github repository

[7]95% CI for all 900 questions = $[93.1, 99.5]$

[8]In traditional Translate-Train, the model is finetuned on translated training inputs for each language *individually*.

5.2 Multilingual Generalization of MLMs and LLMs on BELEBELE

Schematically, the performance of a language model in a given language is related to two key factors. (i) First, the amount of pretraining data in the target language. As predicted by the scaling laws (Kaplan et al., 2020), performance in a language increases monotonically with the amount of pretraining tokens. (ii) Second, the cross-lingual transfer happening between languages in the pretraining data and the target language at inference time (Conneau et al., 2020a,b). This transfer is impacted by a combination of typological, script, and lexical similarities between the pretraining languages and the target language (Muller et al., 2021a, 2023). These two factors are hard to disentangle due to the scale (up to ∼1T tokens) and the potential language leaks of large-scale pretraining corpora (Kreutzer et al., 2022). Thanks to BELEBELE's quality and scale, we provide detailed evidence of both impacting the multilingual generalization of the models.

Impact of Pretraining Language Distribution One of the key differences between the MLMs and LLMs evaluated is their pretraining data distribution and parameter size, explaining the large performance differences between them. For instance, LLAMA 2 largely outperforms XLM-R on high-resource languages, but only achieves accuracy 50 on about half the amount of languages as XLM-R (See Table 2). This difference between the MLMs and LLMs evaluated is illustrated in Fig. 3. However, despite this gap, all LLMs evaluated perform surprisingly well on a large number of languages. For instance, LLAMA-2-CHAT is above 35 accuracy (i.e. 10 above random) for 59 languages. This shows that English-centric LLMs are a promising starting point to build multilingual models.

Machine Translation for Zero-Shot Our Translate-Test evaluations show that using machine translation into English strongly outperforms LLAMA-2-CHAT (70B) performance in the original target language. Across 91 evaluated languages, only 2 are non-trivially better in-language (German and Italian), compared to 68 (none high-resource) for which translating to English is better. none of which are considered high-resource. Compared to LLAMA-2-CHAT having zero-shot accuracy above 50% for 33 languages, it has 71 in Translate-Test (see Appendix A.6.5).

In addition, we evaluate machine-translating the task instructions to the target language. For around

Model	Variant	Eval Setting	AVG	% ≥ 50	% ≥ 70	eng_Latn
Translate-Test (English) on 91 non-English languages in Zero-Shot						
LLAMA-2-CHAT	70B	Translate-Test	**57.1**	**78.0%**	2.2%	78.8
LLAMA-2-CHAT	70B	In-Language	44.1	35.2%	2.2%	78.8
Translated Instructions in 89 non-English languages Zero-Shot						
LLAMA-2-CHAT	70B	In-Language Translated Instructions	38.7	36.0%	**7.9%**	78.8
LLAMA-2-CHAT	70B	English Instructions	**44.9**	**37.1%**	3.4%	78.8

Table 3: Results of LLAMA-2-CHAT in zero-shot in two machine translation-based evaluation settings; Translate-Test (passages, questions, answers translated back to English) and evaluations with the English instructions translated to the target language. The traditional setting on the same languages is provided for comparison.

25 languages, the translated instructions were not well understood (i.e. accuracy less than random), correlating strongly with already low-score languages. For the rest, the performance relative to using English instructions is mixed, though languages that already scored highly had the largest accuracy boost from in-language instructions.

Impact of Sub-Word Tokenization We reaffirm a correlation between increasing vocabulary size and performance on lower resource languages (Liang et al., 2023). XLM-V has a massive 900k-token vocabulary that allocates capacity for each individual language and de-emphasizes token sharing between languages. XLM-V outperforms XLM-R and INFOXLM (250k vocabulary size) on low-resource languages even though they all have the same architecture and are trained on the same dataset (CC-100). GPT3.5-TURBO (100k vocabulary size),[9] FALCON (65k vocabulary size), and LLAMA 2 (32k vocabulary size) all fall off abruptly for medium- and low- resource languages. Larger vocabulary size may explain why FALCON 40B performs equivalent to LLAMA 1 30B despite having been pretrained on fewer non-English tokens.

Scaling effect on Multilingual Generalization We report in Fig. 4 the impact of model sizes on performance on the BELEBELE benchmark across six language families and English. We find that scale is critical for LLAMA to perform reading comprehension as the 7B checkpoint performs slightly above chance in English even. As the parameter size increases, performances across the board increases significantly. Only the 30B and 65B checkpoints are able to perform non-trivially in language families not reported to be in the pretraining corpus (Japanese and Greek). However, unlike other language families such as Romance and Germanic, the performance becomes non-trivial only with the 30B and 65B checkpoints. Results like this suggest that generalizing to distant languages after English-centered pretraining requires more parameters.

Impact of Script Comparing the Romanized versions with the original scripts for Hindi, Urdu, Bengali, Sinhala, Nepali, and Modern Standard Arabic, we find that all models except FALCON perform stronger in the native script than in the Latin script (see Appendix A.6.4). However, the native scripts are allegedly not present in the pretraining data for LLAMA 2 and FALCON. For the Indo-Aryan languages, we hypothesized cross-lingual transfer would be higher in the Latin variant since the tokenization will be more suitable and there is opportunity for shared subwords (anchor points) (Conneau et al., 2020b; Muller et al., 2020; Pfeiffer et al., 2021; Muller et al., 2021a). However, this only seems to be the case for FALCON. The results generally suggest the models were pretrained on significant samples in the native script (perhaps due to code-switching or poor language identification).

6 Conclusion

A fundamental limitation to conducting sound evaluations of the capabilities of language models in low-, or even moderate-, resource languages is the availability of annotated benchmarks. This paper presents a massive dataset, BELEBELE, consisting of passages and multiple-choice questions evaluating reading comprehension in 122 languages. This benchmark enables critical evaluation of reading comprehension capabilities of LLMs in English and top languages. In addition, the dataset is the first of its kind in many medium- and low-resource languages, enabling unprecedented insight into the multilingual capabilities of language models. We present results from a number of famous MLMs and LLMs in different evaluation settings. But given all the evaluations and experiments this dataset enables, we hope BELEBELE unveils further insights that contribute to the development of NLP systems beyond high-resource language.

[9] According to https://github.com/openai/tiktoken.

Limitations

Even with our extensive quality assurance, we warn that "translationese" may cause accuracy on non-English languages to not be directly comparable to English. Often, the *perfect* translation does not exist.

In addition, BELEBELE was designed to measure the reading comprehension abilities of NLP systems across 122 languages. We specifically align as much as possible with translation choices made in the creation of FLORES. Therefore, by-design the samples collected do not capture language- and culture-specific phenomena such as formality (Ersoy et al., 2023), values (Kovač et al., 2023), and aboutness (Hershcovich et al., 2022). Following BELEBELE, building NLP systems inclusive of all cultures and languages will require the release of benchmarks that capture these phenomena.

As briefly mentioned in Section 3.3, annotators discovered a few quality issues with FLORES. Some of them are likely due to style/dialect differences between annotators, but many seem to not be. It's rare enough, thanks to the extensive quality-assurance loops implemented by the NLLB team and the LSP. However, over the scale of 122 languages a fair number of issues have arisen, especially in lower-resource languages. Since updating the FLORES dataset is not in scope for this project, we deliberated on each with the LSP to maximize both appropriateness of the question/answers translations and cross-language consistency.

In general, our model analyses are limited by inconsistent documentation on the composition of pretraining corpora used. To enable further understanding the interaction of multilingual text in training and enable progress in the field, we point to two critical research directions. First, (i) better language identification systems: popular language identification models are trained on a restricted number of languages and domains and only work at the sentence level (Bojanowski et al., 2017), limiting their abilities to track languages in code-switched data and embedded text. Second, (ii) we encourage LLM developers to improve reporting on pretraining language distribution. This is necessary for the research community to understand the cross-lingual transfer capabilities of LLMs and to improve NLP system design for low-resource languages.

Ethics Statement

We removed BLOOMZ (Muennighoff et al., 2023) from our experiment results upon discovery that it was fine-tuned for translation on FLORES-200. However, as alluded to in Section 4.1, we present GPT3.5-TURBO results on BELEBELE even though we cannot verify that FLORES was not in its pretraining or finetuning data. Because of this, comparing results on GPT3.5-TURBO to other models may be unfair given the lack of transparency on training data.

In general, our decision to open-source BELEBELE may compromise future benchmarking as the samples may get collected into large pretraining corpora, undermining fair comparison. This is especially the case for zero- or few-shot evaluation. We determine the value of open-sourcing the full dataset to far outweigh these considerations.

References

Manish Agarwal and Prashanth Mannem. 2011. Automatic gap-fill question generation from text books. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–64, Portland, Oregon. Association for Computational Linguistics.

Kaveri Anuranjana, Vijjini Anvesh Rao, and Radhika Mamidi. 2019. Hindirc: A dataset for reading comprehension in hindi. In *20th International Conference on Computational Linguistics and Intelligent Text*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel R. Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. New protocols

and negative results for textual entailment data collection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8203–8214, Online. Association for Computational Linguistics.

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7422–7435, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In *International Conference of the Italian Association for Artificial Intelligence*.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. 2020. SberQuAD – russian reading comprehension dataset: Description and analysis. In *Lecture Notes in Computer Science*, pages 3–15. Springer International Publishing.

Asım Ersoy, Gerson Vizcarra, Tahsin Mayeesha, and Benjamin Muller. 2023. In what languages are generative language models the most formal? analyzing formality distribution across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2650–2666, Singapore. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Deepak Gupta, Surabhi Kumari, Asif Ekbal, and Pushpak Bhattacharyya. 2018. MMQA: A multi-domain multi-lingual question-answering framework for English and Hindi. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Momchil Hardalov, Ivan Koychev, and Preslav Nakov. 2019. Beyond English-only reading comprehension: Experiments in zero-shot multilingual transfer for Bulgarian. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 447–459, Varna, Bulgaria. INCOMA Ltd.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5427–5444, Online. Association for Computational Linguistics.

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

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Davis Liang, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: Overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13142–13152, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Yash Madhani, Sushane Parthan, Priyanka Bedekar, Gokul Nc, Ruchi Khapra, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Khapra. 2023. Aksharantar: Open Indic-language transliteration datasets and models for the next billion users. In *Findings*

of the Association for Computational Linguistics: EMNLP 2023, pages 40–57, Singapore. Association for Computational Linguistics.

Chaitanya Malaviya, Sudeep Bhatia, and Mark Yatskar. 2022. Cascading biases: Investigating the effect of heuristic annotation strategies on data and models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6525–6540, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 42–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021a. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021b. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Benjamin Muller, Deepanshu Gupta, Jean-Philippe Fauconnier, Siddharth Patwardhan, David Vandyke, and Sachin Agarwal. 2023. Languages you know influence those you learn: Impact of language characteristics on multi-lingual text-to-text transfer. In *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*, volume 203 of *Proceedings of Machine Learning Research*, pages 88–102. PMLR.

Benjamin Muller, Benoît Sagot, and Djamé Seddah. 2020. Can multilingual language models transfer to an unseen dialect? a case study on north african arabizi. *ArXiv*, abs/2005.00318.

Nikita Nangia and Samuel R. Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy. Association for Computational Linguistics.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1221–1235, Online. Association for Computational Linguistics.

Team NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Meta Research*.

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. MCScript2.0: A machine comprehension corpus focused on script events and participants. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 103–117, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Bejan, and Andrew Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium - Technical Report*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. 2022. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Yuan Sun, Sisi Liu, Chaofan Chen, Zhengcuo Dan, and Xiaobing Zhao. 2021. Construction of high-quality Tibetan dataset for machine reading comprehension. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 208–218, Huhhot, China. Chinese Information Processing Society of China.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset

for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Appendix

A.1 Languages and Variants

Language Code	Name in English	Script	Family
acm_Arab	Mesopotamian Arabic	Arab	Afro-Asiatic
afr_Latn	Afrikaans	Latn	Germanic
als_Latn	Tosk Albanian	Latn	Paleo-Balkanic
amh_Ethi	Amharic	Ethi	Afro-Asiatic
apc_Arab	North Levantine Arabic	Arab	Afro-Asiatic
arb_Arab	Modern Standard Arabic	Arab	Afro-Asiatic
arb_Latn	Modern Standard Arabic (Romanized)	Latn	Afro-Asiatic
ars_Arab	Najdi Arabic	Arab	Afro-Asiatic
ary_arab	Moroccan Arabic	Arab	Afro-Asiatic
arz_Arab	Egyptian Arabic	Arab	Afro-Asiatic
asm_Beng	Assamese	Beng	Indo-Aryan
azj_Latn	North Azerbaijani	Latn	Turkic
bam_Latn	Bambara	Latn	Mande
ben_Beng	Bengali	Beng	Indo-Aryan
ben_Latn	Bengali (Romanized)	Latn	Indo-Aryan
bod_Tibt	Standard Tibetan	Tibt	Sino-Tibetan
bul_Cyrl	Bulgarian	Cyrl	Balto-Slavic
cat_Latn	Catalan	Latn	Romance
ceb_Latn	Cebuano	Latn	Austronesian
ces_Latn	Czech	Latn	Balto-Slavic
ckb_Arab	Central Kurdish	Arab	Iranian
dan_Latn	Danish	Latn	Germanic
deu_Latn	German	Latn	Germanic
ell_Grek	Greek	Grek	Hellenic
eng_Latn	English	Latn	Germanic
est_Latn	Estonian	Latn	Uralic
eus_Latn	Basque	Latn	Basque
fin_Latn	Finnish	Latn	Uralic
fra_Latn	French	Latn	Romance
fuv_Latn	Nigerian Fulfulde	Latn	Atlantic-Congo
gaz_Latn	West Central Oromo	Latn	Afro-Asiatic
grn_Latn	Guarani	Latn	Tupian
guj_Gujr	Gujarati	Gujr	Indo-Aryan
hat_Latn	Haitian Creole	Latn	Atlantic-Congo
hau_Latn	Hausa	Latn	Afro-Asiatic
heb_Hebr	Hebrew	Hebr	Afro-Asiatic
hin_Deva	Hindi	Deva	Indo-Aryan
hin_Latn	Hindi (Romanized)	Latn	Indo-Aryan
hrv_Latn	Croatian	Latn	Balto-Slavic
hun_Latn	Hungarian	Latn	Uralic
hye_Armn	Armenian	Armn	Armenian
ibo_Latn	Igbo	Latn	Atlantic-Congo
ilo_Latn	Ilocano	Latn	Austronesian
ind_Latn	Indonesian	Latn	Austronesian
isl_Latn	Icelandic	Latn	Germanic
ita_Latn	Italian	Latn	Romance
jav_Latn	Javanese	Latn	Austronesian
jpn_Jpan	Japanese	Jpan	Japonic
kac_Latn	Jingpho	Latn	Sino-Tibetan
kan_Knda	Kannada	Knda	Dravidian
kat_Geor	Georgian	Geor	kartvelian
kaz_Cyrl	Kazakh	Cyrl	Turkic
kea_Latn	Kabuverdianu	Latn	Portuguese Creole
khk_Cyrl	Halh Mongolian	Cyrl	Mongolic
khm_Khmr	Khmer	Khmr	Austroasiatic
kin_Latn	Kinyarwanda	Latn	Atlantic-Congo
kir_Cyrl	Kyrgyz	Cyrl	Turkic
kor_Hang	Korean	Hang	Koreanic
lao_Laoo	Lao	Laoo	Kra-Dai
lin_Latn	Lingala	Latn	Atlantic-Congo
lit_Latn	Lithuanian	Latn	Balto-Slavic
lug_Latn	Ganda	Latn	Atlantic-Congo
luo_Latn	Luo	Latn	Nilo-Saharan
lvs_Latn	Standard Latvian	Latn	Balto-Slavic

mal_Mlym	Malayalam	Mlym	Dravidian
mar_Deva	Marathi	Deva	Indo-Aryan
mkd_Cyrl	Macedonian	Cyrl	Balto-Slavic
mlt_Latn	Maltese	Latn	Afro-Asiatic
mri_Latn	Maori	Latn	Austronesian
mya_Mymr	Burmese	Mymr	Sino-Tibetan
nld_Latn	Dutch	Latn	Germanic
nob_Latn	Norwegian Bokmål	Latn	Germanic
npi_Deva	Nepali	Deva	Indo-Aryan
npi_Latn	Nepali (Romanized)	Latn	Indo-Aryan
nso_Latn	Northern Sotho	Latn	Atlantic-Congo
nya_Latn	Nyanja	Latn	Afro-Asiatic
ory_Orya	Odia	Orya	Indo-Aryan
pan_Guru	Eastern Panjabi	Guru	Indo-Aryan
pbt_Arab	Southern Pashto	Arab	Indo-Aryan
pes_Arab	Western Persian	Arab	Iranian
plt_Latn	Plateau Malagasy	Latn	Austronesian
pol_Latn	Polish	Latn	Balto-Slavic
por_Latn	Portuguese	Latn	Romance
ron_Latn	Romanian	Latn	Romance
rus_Cyrl	Russian	Cyrl	Balto-Slavic
shn_Mymr	Shan	Mymr	Kra-Dai
sin_Latn	Sinhala (Romanized)	Latn	Indo-Aryan
sin_Sinh	Sinhala	Sinh	Indo-Aryan
slk_Latn	Slovak	Latn	Balto-Slavic
slv_Latn	Slovenian	Latn	Balto-Slavic
sna_Latn	Shona	Latn	Atlantic-Congo
snd_Arab	Sindhi	Arab	Indo-Aryan
som_Latn	Somali	Latn	Afro-Asiatic
sot_Latn	Southern Sotho	Latn	Atlantic-Congo
spa_Latn	Spanish	Latn	Romance
srp_Cyrl	Serbian	Cyrl	Balto-Slavic
ssw_Latn	Swati	Latn	Atlantic-Congo
sun_Latn	Sundanese	Latn	Austronesian
swe_Latn	Swedish	Latn	Germanic
swh_Latn	Swahili	Latn	Atlantic-Congo
tam_Taml	Tamil	Taml	Dravidian
tel_Telu	Telugu	Telu	Dravidian
tgk_Cyrl	Tajik	Cyrl	Iranian
tgl_Latn	Tagalog	Latn	Austronesian
tha_Thai	Thai	Thai	Kra-Dai
tir_Ethi	Tigrinya	Ethi	Afro-Asiatic
tsn_Latn	Tswana	Latn	Atlantic-Congo
tso_Latn	Tsonga	Latn	Afro-Asiatic
tur_Latn	Turkish	Latn	Turkic
ukr_Cyrl	Ukrainian	Cyrl	Balto-Slavic
urd_Arab	Urdu	Arab	Indo-Aryan
urd_Latn	Urdu (Romanized)	Latn	Indo-Aryan
uzn_Latn	Northern Uzbek	Latn	Turkic
vie_Latn	Vietnamese	Latn	Austroasiatic
war_Latn	Waray	Latn	Austronesian
wol_Latn	Wolof	Latn	Atlantic-Congo
xho_Latn	Xhosa	Latn	Atlantic-Congo
yor_Latn	Yoruba	Latn	Atlantic-Congo
zho_Hans	Chinese (Simplified)	Hans	Sino-Tibetan
zho_Hant	Chinese (Traditional)	Hant	Sino-Tibetan
zsm_Latn	Standard Malay	Latn	Austronesian
zul_Latn	Zulu	Latn	Atlantic-Congo

Table 4: The 122 Languages & Scripts in BELEBELE.

As mentioned in Section 3, Bengali, Hindi, Sinhala, Nepali, Urdu, and Modern Standard Arabic are present twice, once in their respective native scripts and once in the Latin script. Chinese is also present twice, in Simplified and Traditional characters. There are 50 Indo-European languages, which we decide to display in smaller language families. Even so, Indo-Aryan is the most common language family (17), followed by Atlantic-Congo (16) and Afro-Asiatic (16).

Note that the FLORES-200 code isn't exactly the same as the older FLORES-101 code, see the FLORES website for details.

A.2 Training Set

To create a training and development set to enable model finetuning to the task for adequate evaluation, we considered a diverse set of multiple-choice question-answering datasets. As there were little options for similarly-formatted datasets outside of English, we only selected English datasets.

After considering 33 different MRC datasets, we determine the most compatible to be RACE (Lai et al., 2017), SCIQ (Welbl et al., 2017), MULTIRC (Khashabi et al., 2018), MCTEST (Richardson et al., 2013), MCSCRIPT2.0 (Ostermann et al., 2019), and RECLOR (Yu et al., 2020). For each of the 6 datasets, we unpack and restructure the passages and questions from their respective formats. Some required more reformatting, such as MCScript2.0, which only has two multiple choice options. For this dataset, we sampled answers from different questions to complete 4 candidate answers. For each, we then filter out less suitable samples, such as questions with multiple correct answers (as in MultiRC), excessively long passages, or fill-in-the-blank questions. We then created subgroups of questions based of surface characteristics such as passage length, question length, and topic. We experiment with these different strata to train the best ROBERTA-base model (Liu et al., 2019) evaluated on the English set from BELEBELE in order to empirically validate the ability to teach a model the correct task. We use these empirical evaluations to finalize the training set, along with a development set of associated samples from the respective validation sets of the 6 above datasets. No test sets were included.

In the end, the dataset comprises 67.5k training samples and 3.7k development samples, more than half of which are from RACE. We provide a script at our github repo to reconstruct this dataset for anyone to perform task finetuning.

Note: Since the training set is a joint sample of other datasets, it is governed by a different license than BELEBELE. Most importantly, we do not claim any of that work or datasets to be our own. See Appendix A.3.

A.3 Licensing

The BELEBELE dataset is licensed under CC-BY-SA, as is the case for the underlying FLORES-200. Please refer to our github repo for more information.

The training set and assembly code is, however, licensed differently. The majority of the training set (data and code) is licensed under CC-BY-NC, therefore the use of BELEBELE for commercial purposes requires a different data solution for model finetuning.

A.4 Experiment Details

A.4.1 Model fine-tuning

As discussed in Section 4.2, for the fine-tuning of MLMs, we use the training set detailed in Appendix A.2. For all settings, the training was performed using the HuggingFace transformers library. We use the development set for hyperparameter search and evaluate the two best training runs on the BELEBELE splits.

For Translate-Train-All, we use machine translation on passages, questions, and answers *separately*. Since there is almost 100x available data for this setting, we limit the training and validation sample to 650k and only train one epoch.

A.4.2 In-Context Learning Prompt

As stated in Section 4.2, for 5-shot in-context learning, examples are sampled from the English training set and prompted to the model. The template used is as follows:

```
P: <passage> \n Q: <question> \n A: <mc answer 1> \n B: <mc answer 2> \n C: <mc answer 3> \n D: <mc answer 4> \n Answer: <Correct answer letter>
```

Within the answers {A, B, C, D}, we determine the prediction to be the one with the highest probability (relative to the others). For all our results, we report the average score over 3 runs.

A.4.3 Zero-Shot Instructions

As stated in Section 4.2, we evaluate both GPT3.5 and LLAMA-2-CHAT in the zero-shot setting by describing the task in natural language. We present the passage, question, and four possible answers, and instruct the model to provide the letter “A”, “B”, “C” or “D ” as the answer. The instructions are given in

English for all languages. We perform post-processing steps and accept answers predicted as e.g. "(A)" instead of "A". The instructions and post-processing code are provided at our github repo.

For the *In-language Translated Instructions* setting, we replicate the above, except present LLAMA-2-CHAT (70B) with instructions that are machine translated to the target language. We do not translate (or transliterate) the lettered answers we ask for ("A", "B", "C" or "D "). Therefore, the post-processing steps are the same.

For the *Translate-Test* setting, we present the natural language instructions in English. The passage, questions, and answers, however, have been machine-translated *individually* back to English.

A.5 Annotation Guidelines

A.5.1 MCQA Annotation Guidelines

The following is an abridged version of the particularized instructions provided to annotators for the task of creating a multiple-choice question-answering dataset. As mentioned, we additionally provided a positive and negative example for each guidelines to the annotators.

1. Ensure that all answers, if not most, are decently plausible to require the test-taker to fully read and understand the passage.

2. In order to make the questions not overly easy, ensure that if the correct answer is word-for-word from the passage, at least some of the wrong answers are as well. This is to ensure that the person answering can't just guess the right answer based off identifying the answer in the passage through skimming.

3. Make the questions as specific as possible, leave nothing to ambiguity.

4. It should not be possible to answer the question without having read the passage, but the question must be answerable using just the passage (no external knowledge). We encourage the use of phrases such as "According to the passage…" in the questions if there may be ambiguity.

5. Try to sometimes have the answers be word-for-word in the passage and for other questions, the answers be in your own words. We would like a balance of these two categories of question.

6. Don't write questions with double negatives that may trivially fool the person answering.

7. Try to ask questions in your own words and don't copy and paste entire phrases from the paragraph. This allows us to truly evaluate the comprehension as opposed to recognizing patterns in the way the question is extracted from the text. That being said, making the questions comprehensible and including details is very important, as mentioned above.

8. Avoid using ambiguous qualifiers (for example, "very" or "really") or having the question centered around an extreme (for example, "the greatest" or "the most"). These phrases may leave room for ambiguity and subjectivity. In addition, qualified statements change with time and may not be valid in the future.

A.5.2 Translation Specifications

To align the dialect or langauge variant to FLORES, we instructed the LSP to use the same localization as in the FLORES-200 creation a few years prior. To align style, formality, and wording, we supplemented the traditional translation guidelines with the following:

> Given that the associated translated passage is already established (and not subject to change as it is in a published dataset), the translations of the questions and answers have to be fully compatible with it. This means that for dates, proper nouns, units of measure, etc. where there is potential ambiguity, the translators have to follow what was done for the passage, even if they disagree that it is the more correct translation.
>
> For example,

Hungarian Translated passage: “Jesus készen áll a vitára…”

English question: “What was Jesus working on when…”

Therefore, in the questions and answers Jesus must be translated as “Jesus” and not “Jézus”.

A.6 Detailed Results Tables

A.6.1 Cross-Lingual MLMs

Full Results for Cross-Lingual MLMs

Evaluation	Finetune in English			Translate-Train-All		
Model Name	XLM-V	INFOXLM	XLM-R	XLM-V	INFOXLM	XLM-R
Size/Variant	large	large	large	large	large	large
AVG	55.6	**56.2**	54.0	**60.2**	60.0	58.9
PCT Above 50	**69.7%**	67.2%	64.8%	**76.2%**	70.5%	69.7%
PCT Above 70	21.9%	**28.9%**	15.7%	33.1%	**37.2%**	36.4%
eng_Latn	76.2	**79.3**	76.2	77.8	**81.2**	78.7
acm_Arab	51.2	57.3	55.4	55.3	57.6	59.2
afr_Latn	69.3	72.7	69.1	72.3	75.1	74.3
als_Latn	68.4	68.9	64.9	70.8	72.2	71.4
amh_Ethi	53.1	52.9	52.6	61.6	60.0	60.7
apc_Arab	56.1	58.8	57.9	57.7	60.6	61.9
arb_Arab	67.2	71.0	69.8	70.6	75.0	74.3
arb_Latn	29.3	32.2	27.6	31.6	33.4	30.6
ars_Arab	55.6	59.9	58.9	61.1	65.8	65.9
ary_Arab	43.8	48.7	44.0	48.0	52.8	52.6
arz_Arab	56.9	60.2	57.6	61.4	64.9	66.1
asm_Beng	53.7	53.6	49.3	58.6	58.8	56.9
azj_Latn	59.7	61.3	59.0	65.0	65.6	65.1
bam_Latn	34.2	34.9	33.2	39.2	39.1	36.9
ben_Beng	60.0	63.4	59.6	65.6	69.6	63.7
ben_Latn	46.8	36.9	38.8	53.0	42.7	48.1
bod_Tibt	24.0	24.9	23.7	24.8	23.3	36.9
bul_Cyrl	72.6	72.0	70.1	74.0	75.3	74.2
cat_Latn	71.6	74.4	72.0	75.7	78.1	74.7
ceb_Latn	45.4	44.1	42.3	52.0	52.6	50.7
ces_Latn	69.9	72.3	69.9	72.3	76.2	74.4
ckb_Arab	29.7	52.3	30.3	36.9	58.0	36.9
dan_Latn	70.8	74.1	72.9	73.0	76.3	74.7
deu_Latn	72.6	75.7	72.9	74.1	78.7	76.7
ell_Grek	70.3	72.3	70.3	73.1	74.9	73.0
est_Latn	63.2	67.2	64.8	68.7	70.7	70.4
eus_Latn	63.6	66.1	64.8	68.2	70.8	70.3
fin_Latn	69.1	72.4	72.2	73.0	75.2	74.9
fra_Latn	73.1	74.2	72.1	74.6	76.8	75.6
fuv_Latn	29.7	27.7	26.4	32.8	30.7	31.1
gaz_Latn	48.8	33.8	36.4	52.6	36.0	43.3
grn_Latn	53.9	37.8	37.9	59.6	40.6	41.9
guj_Gujr	58.7	57.0	54.1	63.3	65.9	63.1
hat_Latn	57.1	39.6	35.2	63.2	44.1	39.8
hau_Latn	51.0	41.1	48.2	53.4	48.1	53.0
heb_Hebr	67.2	68.2	64.8	69.3	72.3	70.6
hin_Deva	57.9	60.2	57.4	63.8	64.3	63.4
hin_Latn	53.1	49.7	46.8	57.6	55.4	58.9
hrv_Latn	70.0	72.4	69.9	71.2	75.3	74.0
hun_Latn	69.7	70.8	70.0	73.1	74.2	72.8
hye_Armn	59.4	61.0	58.9	65.9	66.1	64.7
ibo_Latn	40.1	32.2	31.2	46.8	32.0	32.2
ilo_Latn	37.4	36.3	33.8	38.1	40.6	39.7
ind_Latn	68.9	70.7	68.0	71.3	73.1	70.4
isl_Latn	67.3	66.0	63.8	70.1	68.9	69.0
ita_Latn	70.6	72.8	70.0	71.8	76.4	73.3
jav_Latn	64.2	59.8	60.8	67.2	63.3	66.8
jpn_Jpan	66.4	70.1	67.6	71.3	71.8	71.0
kac_Latn	32.0	29.1	32.1	33.8	34.0	33.3
kan_Knda	61.1	62.0	59.7	66.6	68.4	69.1
kat_Geor	64.7	64.8	63.6	68.0	68.9	67.4

kaz_Cyrl	60.1	61.6	56.8	64.9	65.3	64.7
kea_Latn	44.0	45.2	44.9	48.7	47.7	48.1
khk_Cyrl	56.7	58.8	57.8	61.1	64.6	64.2
khm_Khmr	60.0	59.0	57.7	63.0	64.2	63.8
kin_Latn	35.9	33.6	34.3	39.1	39.1	38.6
kir_Cyrl	65.4	63.4	61.8	68.3	68.2	67.7
kor_Hang	70.1	71.4	68.7	72.9	74.6	74.8
lao_Laoo	55.8	57.6	53.0	63.2	63.6	63.0
lin_Latn	44.7	33.2	30.6	50.9	35.3	34.4
lit_Latn	68.3	69.4	67.2	71.7	72.9	72.0
lug_Latn	39.9	29.4	31.6	47.8	34.7	34.7
luo_Latn	30.3	30.9	30.8	33.7	34.9	33.2
lvs_Latn	70.1	71.3	68.7	74.1	75.6	73.0
mal_Mlym	62.0	65.0	62.7	69.1	68.3	67.1
mar_Deva	62.6	65.2	60.8	69.2	68.8	67.2
mkd_Cyrl	67.8	69.3	65.7	71.0	73.8	72.8
mlt_Latn	37.9	57.1	38.1	40.2	63.7	42.7
mri_Latn	32.0	30.6	32.2	33.0	35.7	34.0
mya_Mymr	56.6	59.1	53.6	62.2	65.1	62.9
nld_Latn	68.4	71.7	71.0	68.6	74.0	72.8
nob_Latn	71.8	73.6	70.7	72.8	75.4	74.2
npi_Deva	58.4	60.7	55.7	64.4	65.8	62.7
npi_Latn	38.3	35.8	33.8	37.4	36.4	34.8
nso_Latn	45.9	31.3	30.0	53.2	34.1	34.7
nya_Latn	31.0	29.2	29.8	34.2	33.0	30.8
ory_Orya	60.8	62.1	58.6	65.6	65.4	63.9
pan_Guru	58.1	59.2	57.8	63.1	62.6	62.0
pbt_Arab	55.4	56.0	51.0	60.6	62.6	61.1
pes_Arab	68.3	69.1	68.2	70.8	73.6	72.0
plt_Latn	55.7	45.6	52.7	61.7	53.4	58.1
pol_Latn	69.0	70.4	67.4	72.1	73.7	72.7
por_Latn	70.9	74.3	70.6	73.8	77.1	74.0
ron_Latn	72.3	72.9	71.3	74.0	76.2	74.8
rus_Cyrl	71.9	73.8	72.2	75.4	76.8	77.1
shn_Mymr	26.9	25.2	26.3	25.0	26.4	27.0
sin_Latn	24.9	34.2	30.7	41.7	38.3	37.3
sin_Sinh	64.4	67.2	62.7	69.8	70.2	68.6
slk_Latn	69.3	71.9	70.2	72.6	76.7	73.0
slv_Latn	69.7	72.2	68.6	71.8	75.4	73.9
sna_Latn	34.8	37.2	33.2	37.1	38.6	35.9
snd_Arab	55.2	56.6	51.9	60.0	61.3	61.3
som_Latn	46.0	39.1	42.6	50.7	46.3	50.7
sot_Latn	46.8	29.3	31.3	52.0	31.9	32.7
spa_Latn	71.0	73.3	71.4	72.7	75.3	76.4
srp_Cyrl	71.0	70.9	71.1	73.6	76.1	75.9
ssw_Latn	39.8	30.6	34.3	47.1	34.3	38.9
sun_Latn	60.9	50.7	55.3	64.2	55.8	59.4
swe_Latn	73.0	75.0	74.2	74.2	76.9	75.1
swh_Latn	64.9	65.3	62.8	69.3	69.2	68.7
tam_Taml	61.8	64.6	61.7	67.4	69.4	65.3
tel_Telu	55.6	57.8	53.6	62.1	63.2	61.1
tgk_Cyrl	38.2	58.6	33.8	39.2	64.3	39.6
tgl_Latn	69.2	67.4	64.7	72.0	70.4	70.0
tha_Thai	63.8	68.1	65.8	69.0	68.9	70.1
tir_Ethi	33.3	36.7	33.8	39.9	42.1	37.7
tsn_Latn	49.0	35.0	30.8	49.8	35.7	34.3
tso_Latn	37.9	36.3	34.2	41.7	39.7	37.1
tur_Latn	66.7	70.2	66.8	70.6	72.0	72.0
ukr_Cyrl	70.4	70.9	71.0	72.3	74.9	75.0
urd_Arab	61.6	63.8	59.3	65.6	68.6	66.3
urd_Latn	42.2	42.6	40.8	49.4	48.9	48.4
uzn_Latn	65.2	66.9	64.4	69.1	70.6	70.2
vie_Latn	69.6	71.1	69.4	73.7	72.9	71.4
war_Latn	46.4	44.7	43.7	47.6	49.3	46.6
wol_Latn	36.8	32.2	30.4	40.6	32.3	32.2
xho_Latn	48.7	36.1	39.0	54.4	40.2	45.4
yor_Latn	35.0	29.3	28.7	38.6	32.0	27.9
zho_Hans	69.8	74.6	71.0	73.7	76.2	74.8
zho_Hant	69.2	72.4	67.1	73.1	74.3	71.3
zsm_Latn	69.1	72.6	69.9	72.4	73.3	72.2

zul_Latn	46.9	36.4	39.0	54.2	39.8	44.1

Table 5: Results of Cross-Lingual MLMs in the two settings described in Section 4.

For all, the *large* version was used which is the same architecture across all three. XLM-V has a significantly larger vocabulary size, leading to more total parameters. We find that in general, INFOXLM and XLM-V are very similar and both out-perform XLM-R across the board. INFOXLM outperforms XLM-V in higher- and medium-resource languages, while XLM-V performs better on the lowest-resource languages. As a result, XLM-V has the most scores above 50, but INFOXLM has more scores above 70.

A.6.2 LLMs

Full Results for Large Language Models

Evaluation	Zero-Shot for Instructed Models		5-shot In-Context Learning			Translate-Train-All
Model Name	GPT3.5-TURBO	LLAMA-2-CHAT	LLAMA 2	LLAMA 1	FALCON	XLM-V
Size/Variant		70B	70B	65B	40B	large
AVG	**50.6**	41.5	**48.0**	40.9	37.3	60.2
PCT Above 50	**43.4**%	27.1%	**38.5**%	25.4%	16.4%	76.2%
PCT Above 70	**28.9**%	2.5%	**26.2**%	12.3%	1.6%	33.1%
eng_Latn	**87.7**	78.8	**90.9**	82.5	77.2	77.8
acm_Arab	51.6	35.9	47.9	37.9	37.6	55.3
afr_Latn	78.3	57.9	75.9	60.7	53.4	72.3
als_Latn	67.1	36.0	45.4	34.9	36.6	70.8
amh_Ethi	28.7	28.9	27.5	27.8	24.8	61.6
apc_Arab	55.6	38.8	51.2	39.6	36.3	57.7
arb_Arab	69.3	42.3	61.7	44.1	38.3	70.6
arb_Latn	31.1	30.2	26.8	28.0	26.3	31.6
ars_Arab	55.1	37.4	50.2	40.7	32.1	61.1
ary_Arab	45.7	32.6	40.6	33.1	32.3	48.0
arz_Arab	56.7	37.3	50.7	37.4	33.0	61.4
asm_Beng	36.0	35.7	32.3	28.9	22.4	58.6
azj_Latn	54.9	33.4	42.2	33.6	34.1	65.0
bam_Latn	31.7	29.4	30.3	28.4	29.7	39.2
ben_Beng	43.6	34.9	39.1	33.4	22.6	65.6
ben_Latn	34.6	30.4	29.6	29.2	32.1	53.0
bod_Tibt	26.6	28.3	25.7	24.9	26.8	24.8
bul_Cyrl	76.0	65.0	80.4	69.3	41.9	74.0
cat_Latn	78.4	68.2	84.6	76.3	58.8	75.7
ceb_Latn	53.3	40.6	50.4	38.9	39.2	52.0
ces_Latn	76.9	65.0	81.1	70.7	65.0	72.3
ckb_Arab	31.8	32.8	28.7	31.6	28.9	36.9
dan_Latn	80.7	66.2	83.6	73.6	56.2	73.0
deu_Latn	83.3	69.4	84.6	76.0	70.1	74.1
ell_Grek	73.0	50.7	64.9	44.2	31.2	73.1
est_Latn	73.1	36.6	53.0	36.3	34.9	68.7
eus_Latn	40.9	31.1	34.7	32.8	38.9	68.2
fin_Latn	77.9	62.7	79.3	55.7	42.8	73.0
fra_Latn	83.1	72.2	86.4	77.5	69.7	74.6
fuv_Latn	26.1	29.8	24.9	25.4	25.1	32.8
gaz_Latn	30.3	29.3	27.8	29.1	24.9	52.6
grn_Latn	34.2	32.2	32.4	30.3	33.8	59.6
guj_Gujr	38.4	31.1	27.1	25.7	24.7	63.3
hat_Latn	51.6	34.1	37.4	33.7	36.2	63.2
hau_Latn	32.2	32.1	28.0	26.4	28.9	53.4
heb_Hebr	64.2	41.4	54.9	41.4	31.1	69.3
hin_Deva	49.1	42.0	52.6	38.4	27.1	63.8
hin_Latn	52.3	39.2	49.0	34.2	40.0	57.6
hrv_Latn	78.4	64.7	79.8	66.9	48.7	71.2
hun_Latn	74.6	61.1	78.8	66.7	37.7	73.1
hye_Armn	35.0	31.9	34.1	32.1	25.4	65.9
ibo_Latn	28.4	30.1	27.4	25.3	30.2	46.8
ilo_Latn	37.1	33.2	36.6	32.1	35.1	38.1
ind_Latn	74.2	61.3	81.4	55.7	52.1	71.3
isl_Latn	62.3	38.0	54.3	42.1	36.4	70.1
ita_Latn	80.0	68.6	84.5	76.1	66.4	71.8

jav_Latn	46.7	37.0	40.3	33.0	36.8	67.2
jpn_Jpan	70.9	56.6	77.6	53.9	49.6	71.3
kac_Latn	30.9	30.7	27.7	28.6	27.8	33.8
kan_Knda	40.6	21.9	25.7	24.4	24.0	66.6
kat_Geor	33.0	34.6	37.8	34.3	23.4	68.0
kaz_Cyrl	35.0	32.4	29.3	32.4	32.6	64.9
kea_Latn	46.0	38.1	45.4	38.1	38.0	48.7
khk_Cyrl	32.0	31.1	29.8	28.4	27.4	61.1
khm_Khmr	30.4	30.6	27.0	28.2	25.0	63.0
kin_Latn	35.2	30.6	29.8	28.5	31.9	39.1
kir_Cyrl	37.9	32.2	34.6	32.5	31.9	68.3
kor_Hang	67.1	56.3	77.8	52.9	40.2	72.9
lao_Laoo	30.0	26.5	24.3	26.2	28.1	63.2
lin_Latn	33.8	31.0	28.0	30.4	29.3	50.9
lit_Latn	72.0	39.7	52.1	39.6	39.3	71.7
lug_Latn	28.4	30.9	29.2	28.3	28.9	47.8
luo_Latn	27.1	31.2	29.4	29.3	29.9	33.7
lvs_Latn	70.8	41.0	51.3	39.0	37.6	74.1
mal_Mlym	34.9	30.1	32.4	30.0	21.2	69.1
mar_Deva	38.3	34.8	41.2	32.9	25.0	69.2
mkd_Cyrl	69.4	55.7	72.5	56.2	38.1	71.0
mlt_Latn	44.8	36.2	44.9	36.7	35.4	40.2
mri_Latn	33.3	31.8	28.5	32.0	29.7	33.0
mya_Mymr	30.3	31.3	24.1	24.2	22.6	62.2
nld_Latn	80.4	66.2	82.2	73.3	66.7	68.6
nob_Latn	79.0	65.7	81.8	70.9	60.8	72.8
npi_Deva	40.4	32.9	40.4	33.0	25.4	64.4
npi_Latn	35.1	30.4	30.2	30.0	30.9	37.4
nso_Latn	33.6	30.1	30.4	27.4	29.3	53.2
nya_Latn	33.2	29.3	27.3	28.7	29.3	34.2
ory_Orya		29.2	24.8	23.9	23.7	65.6
pan_Guru	39.1	33.1	26.3	27.1	23.4	63.1
pbt_Arab	32.3	30.2	30.8	29.4	29.4	60.6
pes_Arab	61.8	41.8	53.9	41.0	35.9	70.8
plt_Latn	32.3	30.5	29.6	31.0	31.4	61.7
pol_Latn	74.7	61.7	79.2	67.0	59.9	72.1
por_Latn	83.0	70.2	86.1	75.4	68.3	73.8
ron_Latn	77.4	65.6	83.4	73.2	66.6	74.0
rus_Cyrl	78.4	67.0	82.7	73.1	48.1	75.4
shn_Mymr		28.2	25.6	22.7	24.0	25.0
sin_Latn	30.4	31.9	33.8	27.9	32.6	41.7
sin_Sinh	32.6	33.4	25.2	29.4	27.7	69.8
slk_Latn	77.3	58.8	75.2	60.4	57.0	72.6
slv_Latn	77.4	62.4	76.7	65.6	43.7	71.8
sna_Latn	35.4	30.2	27.4	28.3	31.6	37.1
snd_Arab	34.1	29.7	30.9	28.9	30.2	60.0
som_Latn	32.4	30.3	27.8	27.6	29.9	50.7
sot_Latn	33.9	30.0	28.9	26.8	29.9	52.0
spa_Latn	79.2	68.4	85.0	74.8	69.2	72.7
srp_Cyrl	74.8	65.1	81.0	70.7	40.2	73.6
ssw_Latn	32.0	30.7	27.7	28.0	30.1	47.1
sun_Latn	38.9	34.9	37.8	30.7	34.1	64.2
swe_Latn	81.7	67.4	82.7	73.7	67.3	74.2
swh_Latn	70.3	35.1	39.6	34.4	36.7	69.3
tam_Taml	32.8	34.4	33.2	31.6	24.4	67.4
tel_Telu	34.6	27.5	25.9	26.6	22.4	62.1
tgk_Cyrl	37.7	32.5	34.0	33.1	32.7	39.2
tgl_Latn	66.7	49.6	68.1	48.3	47.7	72.0
tha_Thai	55.7	38.9	46.2	35.0	33.0	69.0
tir_Ethi	28.4	29.6	24.5	23.5	25.0	39.9
tsn_Latn	31.8	30.1	28.5	24.7	31.2	49.8
tso_Latn	33.4	30.0	30.4	28.0	29.7	41.7
tur_Latn	69.9	47.3	65.4	42.1	39.6	70.6
ukr_Cyrl	72.8	65.7	80.8	69.7	41.9	72.3
urd_Arab	48.3	37.0	43.2	34.7	31.7	65.6
urd_Latn	40.3	34.1	38.0	30.1	34.2	49.4
uzn_Latn	44.1	33.1	35.1	30.6	33.1	69.1
vie_Latn	72.9	59.6	78.4	43.5	41.4	73.7
war_Latn	48.9	39.3	44.4	37.4	38.6	47.6
wol_Latn	29.0	28.9	27.6	26.0	26.8	40.6

xho_Latn	30.0	29.9	28.2	27.6	30.2	54.4
yor_Latn	29.1	30.1	28.3	27.7	27.2	38.6
zho_Hans	77.6	62.4	83.7	64.6	66.0	73.7
zho_Hant	76.3	59.3	82.0	57.7	62.2	73.1
zsm_Latn	74.0	56.4	76.3	51.7	51.3	72.4
zul_Latn	30.4	30.2	29.7	27.1	30.7	54.2

Table 6: Results on LLMs, with comparison to full finetuning on XLM-V

The evaluation settings and models are described in more detail in Section 4. We see that none of these models can understand many of the 122 languages, while demonstrating excellent performance on high-resource languages. The 175B-parameter GPT3.5-TURBO outperforms LLAMA-2-CHAT (70B) across the board and has comparable results to LLAMA 2 (70B) even though it is in zero-shot. Note that GPT3.5-TURBO threw errors when processing characters in Shan (shn_Mymr) and Oriya (ory_Orya) and therefore we could not evaluate the results. For the purposes of aggregated scores, we consider this a score of 25.0. For comparison to fully-finetuned multilingual models, we re-provide the results of XLM-V-large.

A.6.3 Scaling Effect on LLAMA 1

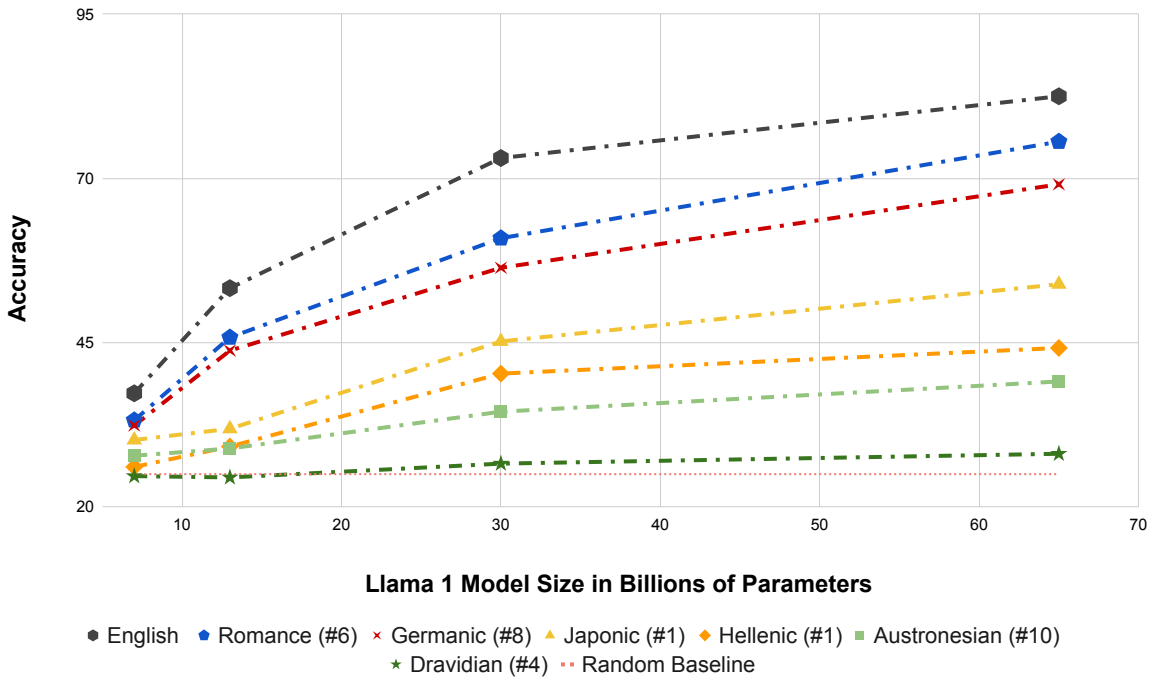

Figure 4: Impact of Models' scale (from 7B to 65B parameters of LLAMA 1) on the performance on BELEBELE for 6 language families and English. The number of languages in a given family is indicated as (#N). LLAMA 1 is evaluated in the 5-shot settings with examples sampled from the training data in English. Scores are average over 3 runs.

A.6.4 Languages in Multiple Scripts

Comparative Results for Languages with Multiple Scripts					
Evaluation	Zero-Shot	Five-Shot		Finetune in English	
Model Name	GPT3.5-TURBO	LLAMA 2 (70B)	FALCON (40B)	INFOXLM-large	AVG
arb_Arab	**69.3**	**61.7**	**38.3**	**71.0**	**60.1**
arb_Latn	31.1	26.8	26.3	32.2	29.1
ben_Beng	**43.6**	**39.1**	22.6	**63.4**	**42.2**

ben_Latn	34.6	29.6	**32.1**	36.9	33.3
hin_Deva	49.1	**52.6**	27.1	**60.2**	47.3
hin_Latn	**52.3**	49.0	**40.0**	49.7	47.8
npi_Deva	**40.4**	**40.4**	25.4	**60.7**	**41.7**
npi_Latn	35.1	30.2	**30.9**	35.8	33.0
sin_Sinh	32.6	25.2	27.7	**67.2**	**38.2**
sin_Latn	30.4	**33.8**	**32.6**	34.2	32.8
urd_Arab	**48.3**	**43.2**	31.7	**63.8**	**46.7**
urd_Latn	40.3	38.0	**34.2**	42.6	38.8
zho_Hant	76.3	82.0	62.2	72.4	73.3
zho_Hans	77.6	83.7	**66.0**	**74.6**	**75.5**

Table 7: Selected results from 3 models in differing settings comparing languages present in multiple scripts.

We find that generally, the performance is higher in the native script than the romanized version, except for FALCON which displays the opposite trend on the 5 Indo-Aryan languages. In Chinese, performance on simplified & traditional are very similar with simplified being higher across all 4. For INFOXLM, we display the English finetuning score.

A.6.5 Translate-Test

Translate-Test Results on 91 languages			
Evaluation Model	**Zero-Shot LLAMA-2-CHAT (70B)**	**Translate-Test, Zero-Shot LLAMA-2-CHAT (70B)**	**Translate-Train-All XLM-V-large**
AVG	44.0	**57.1**	64.9
PCT Above 50	35.2%	**78.0%**	90.1%
PCT Above 70	2.2%	2.2%	42.9%
eng_Latn	87.7	87.7	77.8
fra_Latn	72.2	70.6	73.1
por_Latn	70.2	69.9	70.9
deu_Latn	69.4	65.7	72.6
ita_Latn	68.6	66.1	70.6
spa_Latn	68.4	69.3	71.0
cat_Latn	68.2	67.0	71.6
swe_Latn	67.4	66.1	73.0
rus_Cyrl	67.0	67.3	71.9
dan_Latn	66.2	66.8	70.8
nld_Latn	66.2	67.2	68.4
nob_Latn	65.7	68.3	71.8
ukr_Cyrl	65.7	66.0	70.4
ron_Latn	65.6	67.0	72.3
srp_Cyrl	65.1	66.2	71.0
bul_Cyrl	65.0	67.7	72.6
ces_Latn	65.0	65.6	69.9
hrv_Latn	64.7	65.3	70.0
fin_Latn	62.7	61.1	69.1
slv_Latn	62.4	61.2	69.7
zho_Hans	62.4	71.2	69.8
pol_Latn	61.7	63.0	69.0
ind_Latn	61.3	64.8	68.9
hun_Latn	61.1	62.9	69.7
vie_Latn	59.6	59.4	69.6
zho_Hant	59.3	65.8	69.2
slk_Latn	58.8	66.2	69.3
afr_Latn	57.9	65.0	69.3
jpn_Jpan	56.6	54.8	66.4
zsm_Latn	56.4	67.0	69.1
kor_Hang	56.3	56.7	70.1
mkd_Cyrl	55.7	66.7	67.8
ell_Grek	50.7	67.6	70.3
tgl_Latn	49.6	62.2	69.2
tur_Latn	47.3	62.6	66.7
arb_Arab	42.3	60.7	67.2

hin_Deva	42.0	62.8	57.9
pes_Arab	41.8	59.6	68.3
heb_Hebr	41.4	62.0	67.2
lvs_Latn	41.0	60.9	70.1
ceb_Latn	40.6	62.6	45.4
lit_Latn	39.7	60.8	68.3
hin_Latn	39.2	52.7	53.1
tha_Thai	38.9	54.1	63.8
isl_Latn	38.0	58.1	67.3
jav_Latn	37.0	55.3	64.2
urd_Arab	37.0	59.4	61.6
est_Latn	36.6	59.4	63.2
als_Latn	36.0	63.1	68.4
asm_Beng	35.7	57.7	53.7
swh_Latn	35.1	57.8	64.9
ben_Beng	34.9	61.0	60.0
sun_Latn	34.9	50.8	60.9
mar_Deva	34.8	60.0	62.6
kat_Geor	34.6	57.7	64.7
tam_Taml	34.4	55.9	61.8
urd_Latn	34.1	43.0	42.2
hat_Latn	34.1	56.3	57.1
azj_Latn	33.4	55.6	59.7
sin_Sinh	33.4	57.7	64.4
pan_Guru	33.1	57.6	58.1
npi_Deva	32.9	62.0	58.4
ckb_Arab	32.8	51.3	29.7
kaz_Cyrl	32.4	53.2	60.1
hau_Latn	32.1	43.4	51.0
hye_Armn	31.9	58.0	59.4
mya_Mymr	31.3	46.6	56.6
khk_Cyrl	31.1	52.2	56.7
guj_Gujr	31.1	59.6	58.7
lin_Latn	31.0	40.3	44.7
lug_Latn	30.9	38.7	39.9
ssw_Latn	30.7	43.2	39.8
khm_Khmr	30.6	52.8	60.0
plt_Latn	30.5	46.7	55.7
ben_Latn	30.4	45.1	46.8
som_Latn	30.3	40.8	46.0
pbt_Arab	30.2	48.8	55.4
zul_Latn	30.2	44.4	46.9
nso_Latn	30.1	43.4	45.9
tsn_Latn	30.1	40.4	49.0
yor_Latn	30.1	37.7	35.0
ibo_Latn	30.1	35.3	40.1
mal_Mlym	30.1	63.0	62.0
xho_Latn	29.9	49.2	48.7
fuv_Latn	29.8	29.4	29.7
gaz_Latn	29.3	37.0	48.8
ory_Orya	29.2	57.8	60.8
amh_Ethi	28.9	50.4	53.1
wol_Latn	28.9	39.0	36.8
tel_Telu	27.5	54.3	55.6
lao_Laoo	26.5	47.4	55.8
kan_Knda	21.9	62.0	61.1

Table 8: Comparing LLAMA-2-CHAT Zero-Shot performance In-Language vs Translate-Test on 91 languages, with an aditional comparison to Translate-Train-All finetuning on XLM-V.

We evaluate 91 of the 122 languages in Translate-Test and find that Translate-Test performance dominates in-language performance on a big majority of languages for LLAMA-2-CHAT (70B) in zero-shot. A few head languages such as German and Italian have higher scores in the traditional setting, but nearly all medium-resource languages are better understood with machine translation. For nearly all low-resource languages, the difference is over 20 accuracy points. For comparison, we see that machine translation lifts LLAMA-2-CHAT performance in zero-shot not far from fully finetuned XLM-V-large. This is illustrated more clearly in (Fig. 5) on the next page.

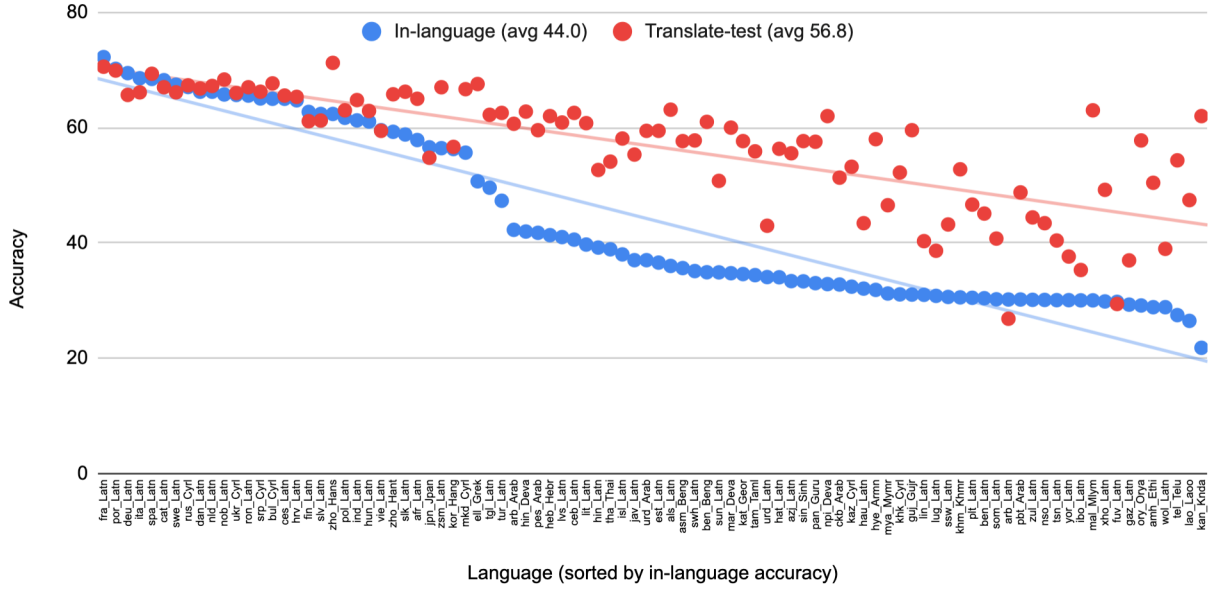

Figure 5: Comparison of LLAMA-2-CHAT (70B) zero-shot performance on Translate-Test and the standard in-language evaluation.