MultiLoKo: a multilingual local knowledge benchmark for LLMs spanning 31 languages

Dieuwke Hupkes* **Nikolay Bogoychev***
Meta
{dieuwkehupkes,nbogoych}@meta.com

Abstract

We present MultiLoKo, a benchmark to evaluate multilinguality in LLMs across 31 languages, with three partitions: a `main` partition containing 500 questions per language, separately sourced for each language to be locally relevant, and two `translated` partitions with human-authored translations from 30 non-English languages to English and vice versa. We also release corresponding machine-authored translations. The data is distributed over two splits: `dev`,s and a blind, out-of-distribution `test` split. MultiLoKo can be used to study a variety of questions regarding the multilinguality of LLMs as well as meta-questions about multilingual benchmark creation. We compute scores for 11 base and chat models and study their average performance and performance parity across languages, how much their ability to answer questions depends on the question language, and which languages are most difficult. *None of the models we studied performs well on MultiLoKo*, as indicated by low average scores as well as large differences between the best and worst scoring languages. We also find a *substantial effect of the question language, indicating suboptimal knowledge transfer between languages*. Lastly, we find that using local vs English-translated data can result in differences of *more than 20 points for the best performing models, drastically changing the estimated difficulty of some languages*. For using machine instead of human translations, we find a weaker effect on ordering of language difficulty, a larger difference in model rankings, and a substantial drop in estimated performance for all models.[2]

1 Introduction

With the growing presence and deployment of LLMs across the world, evaluating their abilities in languages other than English becomes more and more eminent. Yet, studying and evaluating multilinguality in LLMs is a challenging enterprise, and it is hardly exaggerated to call the current state of multilingual evaluation in LLMs insufficient. Older multilingual benchmarks such as XNLI (Conneau et al., 2018) or XCOPA (Ponti et al., 2020) often do not fit the demands for evaluating auto-regressive models and are rarely used for LLM evaluation. Furthermore, their coverage of languages is relatively small compared to the number of languages in which LLMs are intended to be proficient. More often used are benchmarks translated from English, such as MGSM (Shi et al., 2023) or MMMLU (OpenAI, 2025). These benchmarks provide good coverage over many languages, but using translated data comes with its own set of issues. One such issues is that even when human- rather than machine-authored translations are used, translated data is known to differ from native text in several ways (Clark et al., 2020). Furthermore, using translated benchmarks imposes a strong English-centric bias: translated data may be multilingual on the surface, it is not in its content. The benchmarks MLQA (Lewis et al., 2020) and TidyQA (Clark et al., 2020) to some extent address the

*Equal contributions
[2]The data, per-language few-shot examples, evaluation scripts, and prompts can be found in our online.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

issue by sourcing data separately for different languages. Even in their sourcing protocols, however, there is no explicit focus on selecting locally relevant content for the chosen languages. In addition to that, their coverage is again small compared to the above mentioned translated benchmarks.

In response to these issues, we introduce a wide-coverage multilingual benchmark with locally-sourced questions for 31 different languages. Because the benchmark targets multilingual local knowledge, we dub it `MultiLoKo`. The release of MultiLoKo serves two interconnected goals:

1) Provide a better means to evaluate multilinguality in LLMs;
2) Provide data to study the effect of various design choices in multilingual evaluation.

To address our first goal, we create 500 questions per language, written from scratch for each language, using a sourcing protocol specifically designed to ensure local relevance of the question topics. To also reap the benefits of parallel data, we commission both human and machine-authored translations for all non-English questions into English and vice versa, providing a total of 15500 parallel questions, sourced across the 31 languages in the benchmark. The translated data facilitates the study of transfer between languages and also serves our second goal. By comparing the English-translated data with the locally sourced data, we can explicitly compare the adequacy of using translated benchmarks; by comparing human- with machine-authored translations, we can better estimate the potential issues of the latter. To prevent quick overfitting and inadvertent contamination, we release a development set of the benchmark, while test scores can only be obtained through an external provider.We compute average performance and language parity scores on the locally sourced data for 11 models marketed for their multilinguality (§ 5.1); we investigate whether models exhibit knowledge transfer between different languages (§ 5.2); we study the impact of local sourcing versus translating on model rankings and language difficulty (§ 5.3.1); we analyse the difficulty of the included languages through various lenses (Appendix D); and we conduct an analysis into the difference between human- and machine-authored translation (Appendix E). We find that the best performing model is Gemini 2.0 Flash, with an average performance of 34.4 points, and an almost 35 point gap between the best and the worst language. Llama 3.1 405B and GPT4-o are close contenders in terms of average scores (34.3 and 34.0, respectively), but both have substantially higher language gaps (39 and 49 points). Almost across the board, model performances are better when questions are asked in the language to which the content is relevant, indicating suboptimal knowledge transfer between languages, a result that is mirrored by low response-consistency across question language.

Next, we study the relevance of using locally sourced data as opposed to translated English data as well as whether it matters if translations are authored by humans or machines. We find that the estimated difficulty of some languages changes drastically across the two sourcing setups, within the range of 15 points decrease and 8 points increase on average across models. The rank correlation between average language difficulty score is 0.78. Furthermore, individual model scores between local and English-translated data can differ up to 22 points for some languages. However, changing the sourcing setup does not impact model rankings, suggesting that using translated data may be suitable for comparing models but less for model development or language prioritisation. For using machine- instead of human-authored translations, as well, the effect on model ranking is limited (R=0.97), but the difficulty estimates of various languages changes with up to 12 points. Furthermore, using machine translated data results in lower average scores for all models, with drops ranging from 2 to 34% of the human-translated scores.

Outline In the remainder of this paper, we first describe our dataset collection protocol (§ 2) the dataset itself in (§ 3), and our experimental setup (§ 4). In § 5, we present a range of different results. We conclude in § 6 and discuss limitations in Appendix F. Beyond the related work discussed above, we include a discussion of a wider range of multilingual datasets in Appendix A.

2 Dataset collection

Similar to the protocol used by the well-known benchmark SQuAD (Rajpurkar et al., 2016), we source articles from Wikipedia about which we ask annotators to generate questions. After that, we run several rounds of quality control on the generated questions and commission human- and machine-authored translations of all data. Our collection protocol consists of five steps.

Step 1: Paragraph selection We start by sampling the 6K most visited Wikipedia pages for each language for the period of 2016-2021. We then sample paragraphs from those pages by randomly selecting a word in the page and expanding left and right until we reach 3K characters. Next, we ask annotators to judge the local relevance of the samples on a scale from 1 to 5, where 1 refers to topics specific to the language (e.g. a Swedish singer not known outside of Sweden) and 5 to globally well-known topics (e.g. 'Youtube'). We disregard all topics that have a locality score above 3. The full rubric and annotation instructions can be found in Appendix I.1.

Step 2: Question generation In step 2, we ask native speakers to generate challenging questions about the content in the paragraphs. To facilitate automatic scoring, we ask that the questions are closed-form questions, with only one correct short answer. To ensure that the annotation instructions are understandable and appropriate for each locale and the questions of high quality, we run a pilot with 50 questions separately for each language. After our pilot, we commission 500 additional samples for each language, to leave a 10% margin to disregard questions in the rest of the process.

Step 3: Question review For each generated question, we ask a new set of annotators from a separate provider to judge whether the generated questions abide by the annotation instructions, to flag any possible issues, and to mark if the question is useable as is, would be useable with a small adaptation or should be disregarded. We ask annotators to fix small annotation errors on the spot, and as respective vendors that questions with larger issues are replaced.

Step 4: Question answering As a last quality control step, we ask two annotators different from the creator of the question to answer the questions. In this stage, we do not ask annotators to correct questions, but we simply disregard all questions for which either annotator thinks the original answer was incorrect, or the annotator provided an answer not matching the original answer because of ambiguities in the question. The only corrections we allow in this stage are additions of additional, semantically equivalent, correct answers (e.g. 'four' as an alternative to '4').

Step 5: Translation Lastly, we translate the non-English data back to English and vice versa. This allows to study generalisation of knowledge and skills between English and non-English languages and facilitates inspection of the topics and questions for all languages of the dataset, without understanding those languages. We commission both human and machine translations[3] and study their difference as part of our analysis.

3 MultiLoKo the dataset

MultiLoKo consists of three main components: i) the collected data; ii) a set of multilingual prompts to prompt base- and chat models; and iii) a set of metrics.

3.1 The collected data

The data in MultiLoKo consists of several *partitions* and two *splits*.

Partitions MultiLoKo has one `main` partition, containing locally-soured data for 31 languages, and four `translated` partitions. Two of the latter are *human*-translated: `human-translated-from-english`, with translations of English data into the 30 other languages in MultiLoKo and `human-translated-to-english`, containing translations of the non-English subsets into English. The other two are *machine*-translated partitions following the same pattern. All partitions contain 500 samples per language – thus in total 15500 samples in the `main` partition, and 15000 samples in all `translated` partitions. Further statistics about the dataset, such as the distribution over answer types and the average prompt length, can be found in in Appendix B.

Splits Each partition is divided equally over two *splits* containing 250 samples per language: a `dev` split that can be used for development, and a blind `test` split. Until the `test` split is publicly released, results can only be obtained through model submissions.[4] The splits are not random, but constructed such that for each language the most frequently visited pages are in the `dev` split while the least frequently visited pages are in the `test` split, roughly preserving the distribution of answer types (e.g. number, name, year, etc). The `test` split can thus be seen as an out-of-distribution (ood) split, specifically meant to assess generalisation (which is challenging in the context of LLMs, see e.g. Hupkes et al., 2023). In § 5.3.2, we provide an analysis of the extent to which the split is truly an ood split, by analysing its difficulty. The results reported in the results section of the paper are `dev` results.

[3]For the machine translations, we use the Google Translate sentence based cloud API.

[4]More details can be found on `https://github.com/facebookresearch/multiloko/`.

3.2 Prompts and few-shot examples

In the spirit of getting truly multilingually appropriate results, we design the prompts required to run separately for each language and release them along with the data. The prompts are written by a different linguistic experts for each language, in consultation with the benchmark creators to ensure they are appropriate for LLMs. We provide prompts for base models and chat models that allow for incorporating up to five few-shot examples all of which we provide in our github repository.

3.3 Metrics

MultiLoKo has two main metrics and two auxiliary metrics. The two main metrics – *Exact Match accuracy* (EM) and *Gap* – capture the overall performance of MultiLoKo and are computed on the `main` partition, whereas the two auxiliary metrics – *Mother Tongue Effect* (MTE) and *Locality Effect* (LE) – combine information from different partitions. We provide a cheat-sheet in Table 2.

EM and Gap EM indicates the performance of a model on a single language or averaged across languages, as measured by the percentage of times the post-processed model answer verbatim matches one of the answers in the reference list. Gap, defined as the difference between the best and the worst performing language in the benchmark, is a measure of *parity* across the individual languages within the benchmark. Taken together, EM and Gap provide a good indication of how well a model is faring on MultiLoKo. Because both gap and EM are binary metrics that may be open to false negatives, we also considered the partial match metrics BLEU (Papineni et al., 2002), ChrF (Popović, 2015) and *contains*, but we did not find any different patterns using those metrics.

MTE Because of the 2x2 design of MultiLoKo, in which we translated non-English data back to English and vice versa, we can compute several metrics related to locality of the requested information. MTE is one of such metrics. It expresses the impact of asking a question in a language to which that question is relevant. We quantify MTE (for non-English languages only), as the delta between the EM score of the locally sourced data asked in the corresponding language (e.g. asking a question about a local Bengali radio station in Bengali) and the EM score when the same questions are asked in English. A positive MTE indicates that information is more readily available when it is relevant to the language in which it was asked, whereas a negative MTE indicates that the information is more easily accessible in English. MTE is a measure related to transfer as well as language proficiency.

LE The locality effect (LE) is a measure of how much performance on knowledge tasks is over- or underestimated through the use of using translated English data, as opposed to locally relevant data. We quantify the locality effect as the difference in EM for English translated data and locally sourced data. If for a language the English translated data has as a higher EM, the LE is *positive*, indicating that using English translated data likely *overestimating* a model's ability on providing knowledge for that language. If the LE is *negative* the English translated data may provide an *underestimation* of the score for that language. Note that because we often observe both positive and negative LEs for the 30 non-English languages in MultiLoKo, the average LE across languages may be small, even if the differences for individual languages may be large.

4 Experimental setup

We test and showcase our benchmark by running experiments with 11 different models of varying sizes, that were all marketed to have multilingual abilities.

4.1 Models

To test the extent to which MultiLoKo provides useful signal across training stages, we consider both base and chat models. The base models we include in our experiments are Llama 3.1 70B and 405B (Dubey et al., 2024), Mixtral 8x22B (team, 2024), and Qwen 2.5 72B (Qwen et al., 2025), the seven chat models are Gemini 2.0 Flash (Google DeepMind, 2024), GPT4-o (OpenAI et al., 2024), Claude 3.5 Sonnet (Anthropic, 2025), Llama 3.1 70B and 405B Chat, Mixtral 8x22B-it, and Qwen 2.5 72B instruct. As mentioned before, we run chat and base models with separate prompts.

Table 1: **Aggregate results dev.** We report average EM, gap, mother tongue effect and locality effect for all 11 models on the MultiLoKo dev split. For EM, MTE and LE, we also indicate a confidence interval equal to two times the standard error across languages. Models are sorted by average EM.

Model	EM	Gap	Mother tongue effect	Locality effect
Gemini 2.0 Flash	34.39± 2.90	34.80	6.12± 1.90	0.36± 3.40
Llama 3.1 405B	34.31± 2.70	39.20	6.37± 1.70	0.62± 2.70
GPT4-o	33.97± 3.60	48.80	3.08± 2.00	0.35± 2.90
Llama 3.1 405B Chat	27.70± 3.20	40.80	3.97± 2.20	−1.11± 2.70
Llama 3.1 70B	26.92± 2.60	28.80	2.72± 1.70	−0.30± 3.10
Claude 3.5 Sonnet	26.89± 4.40	47.60	N/A	0.81± 2.90
Llama 3.1 70B Chat	21.65± 2.80	42.40	0.49± 1.60	−3.32± 3.30
Mixtral 8x22B	21.64± 4.20	43.60	−2.18± 3.00	−0.65± 2.60
Qwen2.5 72B	19.66± 2.30	28.40	2.45± 2.10	−2.28± 2.70
Mixtral 8x22B-it	10.10± 3.10	39.20	−5.41± 2.00	−0.54± 1.70
Qwen2.5 72B instruct	2.54± 0.70	8.00	−1.52± 1.00	0.43± 0.70

4.2 Experimental setup

We run all experiments with generation temperature set to 0. To facilitate automatic evaluation, we include an instruction to answer questions curtly and precisely, producing only a number/name/etc. Full template information can be found in our github repository. We use a 5-shot prompt for base models and a 0-shot prompt for chat-models. For base models, minimal postprocessing is needed: we lowercase the output, strip punctuation and whitespace, and evaluate the first line. Chat models often deviate from the required format, in various ways that we discuss in Appendix G. To evaluate such models beyond their instruction-following issues, we perform more complex post-processing, aiming to remove any words resembling “answer” from the LLM output, as well as several special cases for English and Japanese. We provide full details about post-processing in Appendix H.

5 Results

We report average model results (§ 5.1), study transfer between languages (§ 5.2) and look in more detail at the dataset itself through the lens of model results (§ 5.3). We report language specific results and differences between using human and machine translated data in Appendices D and E.

5.1 Aggregate results: EM and language gap

In Table 1, we report per-model average EM, the gap between the best and worst language, and average MTE and LE, which we will discuss in a later section. We report average MTE, EM and LE along with a confidence interval equal to two times the standard error across languages, roughly equalling previously used 95% confidence intervals (Madaan et al., 2024; Dubey et al., 2024).

Model performance (EM) In Figure 1 (left), we show a boxplot of the EM scores across models. The best performing models are Gemini 2.0 Flash, Llama 3.1 405B, and GPT4-o, while Mixtral 8x22B and Qwen2.5 72B populate the lower rankings. Somewhat surprisingly, base models are generally outperforming chat models on the benchmark, this is partly due to false refusals and poor instruction following in the chat models. In some cases, however, the chat models simply just provide a qualitatively different answer than the base models. The figure shows that MultiLoKo is a relatively difficult benchmark across the board: the average EM of even the best performing model barely exceeds 30, while the bottom performing models have scores lower than 20. Also EM for the easiest languages (see also Appendix D) remain below 50. Furthermore, for virtually all models performance varies starkly between languages, suggesting that none of the models we considered are evenly multilingual across the 31 languages covered.

Gap While average EM score provides some information about a model’s multilingual abilities, the same EM score can hide many different patterns regarding individual language scores. As we appreciate it is not always practical to consider 31 separate EM scores in model development, we add a second summary metric to the main metrics of MultiLoKo: the gap between the best and worst performing languages, rereresentative of the extent to which a model has achieved parity across languages. Earlier, we already saw that the per-language scores have quite a range for all models. In

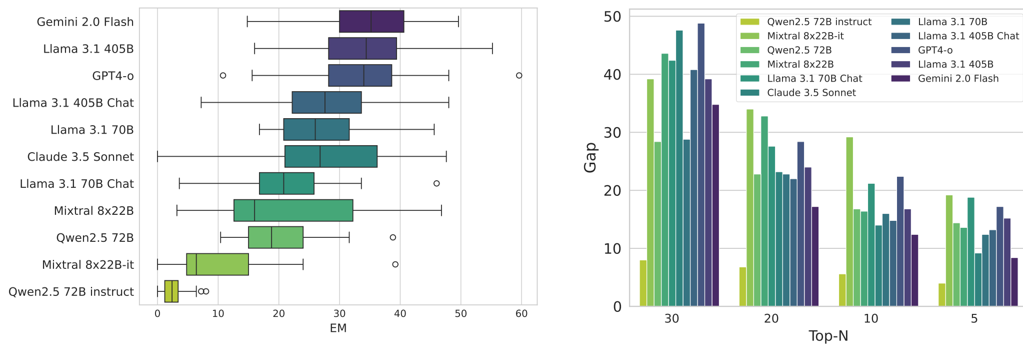


Figure 1: **EM distributions and Gap dev.** Left: Boxplot of EM scores across models, sorted by mean. Right: Difference between the best EM and the worst of the N next best EM scores, per model.

Figure 1 (right), we study this in more detail, by considering the gap between the best language and the next N best language (30 corresponds to the full benchmark). On the right end of the plot, we see that already considering only 5 languages besides English, even the best perform has a gap of over five points – which is relatively large in absolute terms and very large in relative ones – between English and the worst of the remaining languages. For the second best two models, the top-5 gap even exceeds 10 points. As we include more languages, up to the full benchmark, the gap increases, with GPT4-0 showing gap of almost 50 points. The only models for which the gap is small are the models that have overall low performance and thus little space to drop from English, illustrating how gap and average EM provide complementary information about multilingual performance.

5.2 Generalisation across languages

Next, we study whether knowledge generalises across languages.

The mother tongue effect (MTE) First, we compare the EM of models when questions are asked in the language for which the questions were originally sourced with performance when the same questions are asked in English. We quantify this effect with the metric MTE, which expresses the difference in performance between these two settings (see § 3.3). In Figure 2 (left), we show MTE per language, averaged across models.[5] For most languages, performance is higher when the question is asked in the language for which the question is locally relevant. With the exception of Hindi, the languages for which MTE is negative or close to 0 are all languages that perform very poorly also in the mother tongue and for which there is therefore little room for further decrease. From one perspective, the improvements when questions are asked in the low-resource but native languages can be seen as surprising: as models perform much better in English than non-English languages, one may expect performances to go up as a consequence of that. On the other hand, similar 'mother tongue effects' have been observed in earlier studies. For example, Ohmer et al. (2024) found that models are comparatively better at answering factual questions about topics when they are asked in a language to which culture the fact pertains. It appears that also in our case, the effect of accessibility of information in a relevant language wins out over the generally stronger English performance, pointing to a gap in models' ability to generalise knowledge from one language to another. In Figure 2 (right), we further consider the distribution of MTE scores for the top-3 models. Interestingly, this distribution is quite different between models. Despite having comparable average scores, the top-3 performing models differ in their MTE distributions across languages. Of the three models, GPT4-o has the smallest average effect (3.2); Llama 3.1 405B has a much higher average effect (6.6), but less probability mass on the more extreme ranges of the spectrum (min max values of [-7, +12] vs [-9, +13]) Gemini 2.0 Flash is in the middle in terms of average (6.3), but shows the largest variation across languages ([-10, +16]). Note, however, that without studying the actual training data of the models, it is possible to infer that the models have relatively poor transfer across languages, but not conclusively say that one model is better than another: it is also possible that the information sourced for languages with better MTEs was simply better represented in the English data of a model.

[5] Claude 3.5 Sonnet scores were very low on English because of poor instruction following (s see Appendix G). As this is unrelated to lack of transfer or knowledge, we exclude Claude 3.5 Sonnet from all transfer results.

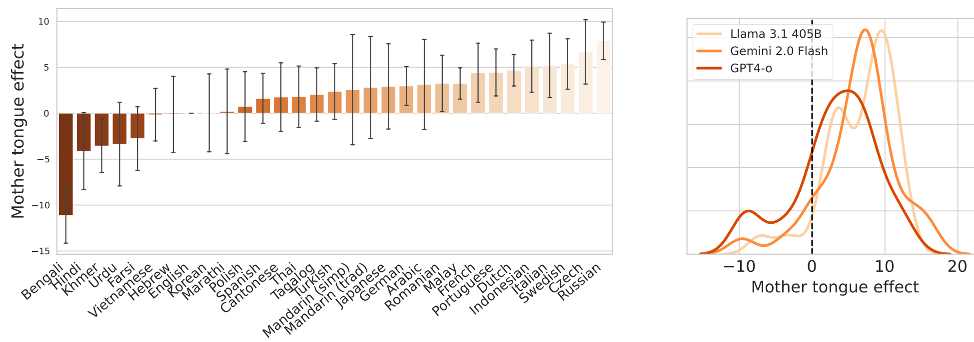


Figure 2: **Mother tongue effect.** Left: Per language MTE, indicating the delta between asking questions in mother tongue vs English. Error bars indicate 2x SE across all models but Claude 3.5 Sonnet. Right: KDE plot of the distribution of MTE scores for the top-3 performing models.

Model	Consistency
Gemini 2.0 Flash	0.46± 0.04
Llama 3.1 405B	0.46± 0.04
Llama 3.1 70B	0.45± 0.03
GPT4-o	0.45± 0.05
Llama 3.1 405B Chat	0.42± 0.04
Qwen2.5 72B	0.40± 0.04
Llama 3.1 70B Chat	0.40± 0.04
Mixtral 8x22B	0.36± 0.05
Mixtral 8x22B-it	0.21± 0.05
Qwen2.5 72B instruct	0.08± 0.03

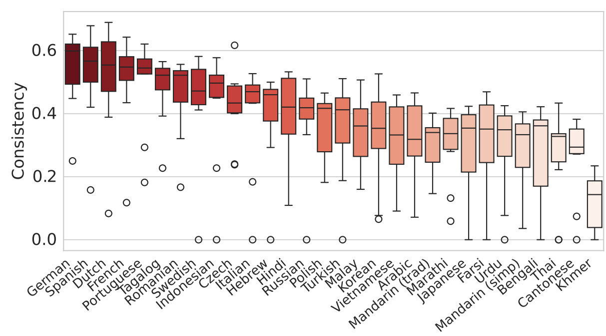


Figure 3: **Consistency results.** Left: Average per-model consistency scores, $\pm$ 2 times the standard error across languages. Right: Boxplot of model consistency scores per language, indicating the relative overlap of correctly answered questions when asked in the mother tongue vs in English.

Consistency across responses Another way to study transfer between languages is to look at the consistency of responses across languages (Qi et al., 2023; Ohmer et al., 2023, i.a.). After all, it is possible for a model with an EM of 30 on both native and translated data to be completely misaligned on *which* questions they respond to correctly. Studying consistency across responses can therefore be seen as a more direct way of studying whether knowledge is equally accessible across languages. In the dataset used by Ohmer et al. (2023), the correct answers are identical across languages, while Qi et al. (2023) use a ranking approach. Neither of their metrics can be directly applied in our case. Rather, we opt for a simpler consistency metric, which quantifies what percentage of the questions that are answered correctly in *either* language are answered correctly in *both* languages. In Figure 3 (left), we show the average consistency of all models; we also show the per-language consistency results in Figure 3 (right). The results confirm our earlier conclusion that much improvements can be made when it comes to knowledge transfer between languages: even for the best performing models, there is an overlap of not even 50% between the questions correctly answered across languages.

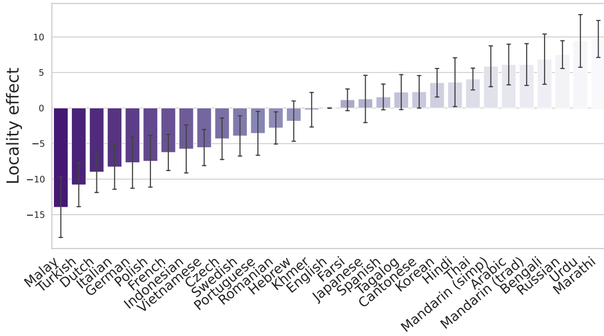
5.3 The dataset

Lastly, we discuss two aspects related to the creation of the dataset.

5.3.1 Locally-sourced vs translated-from-English data

To study the impact of using locally sourced data, we consider the difference between per-language EM on locally sourced data and translated from English data.

Language difficulty First, we look at per-language differences between locally sourced and translated English data. We quantify this with a metric we call the Locality Effect (LE), which tells us how much the estimate of a model's strength in a particular language would have been off if we had chosen to use a translated benchmark rather than a locally sourced one. We plot this difference



Model	Rank correlation language difficulty
Gemini 2.0 Flash	0.54
Llama 3.1 405B	0.65
GPT4-o	0.64
Llama 3.1 405B Chat	0.70
Llama 3.1 70B	0.60
Claude 3.5 Sonnet	0.84
Llama 3.1 70B Chat	0.68
Mixtral 8x22B	0.86
Qwen2.5 72B	0.45
Mixtral 8x22B-it	0.88
Qwen2.5 72B instruct	0.55

Figure 4: **Locality Effect.** Left: Per language LE, expressing the delta in EM between locally sourced and translated English data. Right: Per-model rank correlation between language difficulty of languages on locally sourced vs English translated data.

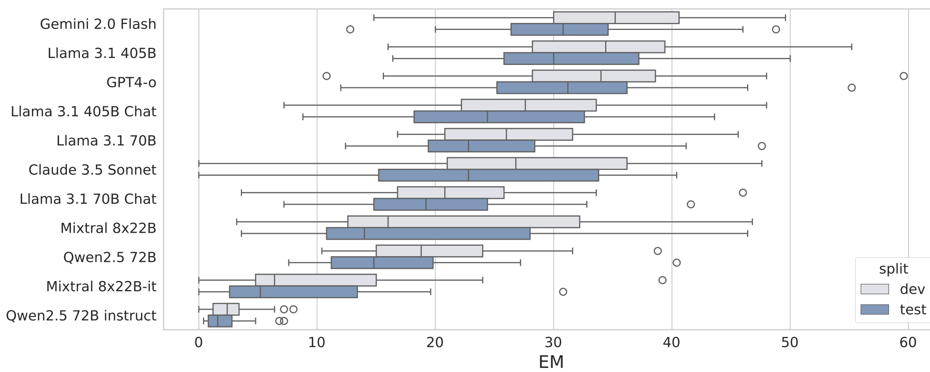


Figure 5: **Average EM, dev versus `test`.** Score distributions of the `dev` (upper bars) and `test` (lower bars) sets. The results show that the `test` set is indeed out of distribution with respect to the `dev` set.

in § 5.3.1 (left). As we can see, the scores between locally and translated English-sourced data can differ quite drastically, almost 15 percentage points averaged across models. For individual models, the differences are even larger. For Llama 3.1 405B, LE ranges from -13 to +17; for Gemini 2.0 Flash from -21 to +15; and for GPT4-o from -22 to +14. The differences are not just in absolute scores; also the ordering of language by difficulty is quite different across the two data collection setups, as can be seen by the per-model rank correlations of language difficulty between the two conditions, shown in § 5.3.1 (right). Using English-translated rather than locally sourced data does thus not only provide different estimates, but may suggest different languages to focus on for improvement.

Model rankings Next, we consider the ranking of the models under the two different data regimes. Interestingly, given the transfer effect, changing from locally to English translated data does not make any difference in the ranking. Also in terms of absolute scores, the difference between the two data collection setups is relatively minor. At least for our type of data, it thus appears that using translated data as opposed to locally sourced data may be a reasonable setup for comparing models on average, though not for getting adequate per-language or set language prioritisation.

5.3.2 The dataset split

As mentioned in the dataset construction, we took the deliberate decision to generate a split based on topic frequency, rather than creating a random split. The aim of this out-of-distribution split is to test generalisation to topics that are more in the tail of the distribution, as well as encourage improvements in multilinguality beyond having a higher score on the specific released MultiLoKo dev set. Of course, however, because of our sourcing method, *all* the topics in MultiLoKo are topics on which information is available on Wikipedia. As training data, Wikipedia is often packaged as a single scrape, this may render our deliberate splitting efforts futile: the fact that a page is less visited does not make it less likely that the specific page is included in the training data. Now, we test if the `dev` and `test` split are in fact distributionally different.

In Figure 5, we show boxplots of `dev` and `test` EM scores for all models under consideration. The plot confirms that the split is indeed to be considered an OOD split: for virtually much all models, the test scores are lower than the dev scores. Across all models, the average dev score is 24, whereas the average test score is 21. This suggests that our test set does indeed contain more tail knowledge than the dev set, despite the aforementioned arguments regarding Wikipedia. Interestingly, this implies that Wikipedia may not be the primary source from which models learn this information.

The difference in difficulty also has bearing on the other metrics: the *gap* between the best and worst performing language) is 37 for `dev` vs 34 for `test`, suggesting that more difficult dat may to some extent hide differences between languages and therefore exemplifying the utility of considering parity along with overall performance. The mother tongue effect, on the other hand, is comparable across `dev` and `test` (1.61 vs 1.56, respectively). For the locality effect, the effect is less interpretable. While the average difference is substantial (-0.6 `dev` vs -1.9 `test`), there is no clear pattern discernable across languages: for some, the effect reduces, whereas for others it increases.

6 Conclusion

Notwithstanding the increasing multinational deployment of LLMs in many parts of the world, adequately evaluating their multilinguality remains a challenging enterprise. Only in part is this due to the scarcity of high-quality and broad-coverage multilingual benchmarks for LLM: perhaps a more pressing issue is that the benchmarks that *are* frequently used for multilingual evaluation virtually all consist of translated English data. While using completely parallel data has its advantages, using translated English data imposes an English-centric bias on the content of the benchmarks, implying that even if the benchmark evaluates multilinguality on the surface, it does not in content. In our work, we aim to address this by presenting `MultiLoKo`, a multilingual benchmark spanning 31 languages that combines the best of both worlds. MultiLoKo contains 500 questions targetting locally relevant knowledge for 31 languages, separately sourced for each language with a protocol specifically designed to ensure local relevance of the question topics. It is also fully parallel, because it contains human-authored translations of the non-English partitions into English and vice versa. As such, it allows to study various questions related to multilinguality, transfer and multilingual benchmark creation. To prevent quick overfitting and inadvertent contamination, we release a development set of the benchmark, while the test set of the benchmarks remains private, at least for the near future.

We use MultiLoKo to analyse 4 base and 7 chat models marketed to be multilingual. We find that the best performing model is Gemini 2.0 Flash, with an average performance of 34.4 points, and an almost 35 point gap between the best and the worst language, followed by Llama 3.1 405B and GPT4-o, which are close contenders in terms of average performance but both have substantially higher language gaps (39 and 49 points). Generally, scores are better when questions are asked in the language to which they are relevant, indicating suboptimal knowledge transfer between languages, a result that is mirrored by low per-sample consistency across question language.

On a meta-level, we study the relevance of using locally sourced data as opposed to translated English data as well as whether it matters if translations are machine- or human-authored. We find that the estimated difficulty of some languages changes drastically across the two sourcing setups, within the range of 15 points decrease and 8 points increase on average. The rank correlation between average language difficulty score is 0.78, and individual model scores can differ up to 22 points for some languages. However, changing the sourcing setup does not impact model rankings, suggesting that using translated data may be suitable for comparing models but less for model development or language prioritisation. For using machine- instead of human-authored translations, as well, the effect on model ranking is limited (R=0.97), but the difficulty estimates of various languages changes with up to 12 points. Furthermore, using machine translated data results in lower average scores for all models, with drops ranging from 2 to 34% of the human-translated scores.

While our results section is extensive already, there are still several parts of MultiLoKo that we did not explore. For instance, because of the sourcing strategy, each native question is coupled with a paragraph that contains the answer to the question. MultiLoKo could thus be transformed into a reading-comprehension benchmark, and we consider studying the difference between the knowledge and reading comprehension setup an interesting direction for future work. Furthermore, each question contains an elaborate long answer intended to explain the short answer. We have not used the long answers in any of our experiments, but foresee interesting directions including studies into CoT prompting or studying answer rationales.

References

Fakhraddin Alwajih, Gagan Bhatia, and Muhammad Abdul-Mageed. Dallah: A dialect-aware multimodal large language model for Arabic. In Nizar Habash, Houda Bouamor, Ramy Eskander, Nadi Tomeh, Ibrahim Abu Farha, Ahmed Abdelali, Samia Touileb, Injy Hamed, Yaser Onaizan, Bashar Alhafni, Wissam Antoun, Salam Khalifa, Hatem Haddad, Imed Zitouni, Badr AlKhamissi, Rawan Almatham, and Khalil Mrini, editors, *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 320–336, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.arabicnlp-1.27. URL https://aclanthology.org/2024.arabicnlp-1.27/.

Anthropic. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2025. Accessed: 2025-04-11.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL https://aclanthology.org/2020.acl-main.421/.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. Is it good data for multilingual instruction tuning or just bad multilingual evaluation for large language models? *CoRR*, abs/2406.12822, 2024. URL https://doi.org/10.48550/arXiv.2406.12822.

Zhihong Chen, Shuo Yan, Juhao Liang, Feng Jiang, Xiangbo Wu, Fei Yu, Guiming Hardy Chen, Junying Chen, Hongbo Zhang, Li Jianquan, et al. Multilingualsift: Multilingual supervised instruction fine-tuning, 2023. URL https://arxiv.org/pdf/2412.15115.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8: 454–470, 2020. doi: 10.1162/tacl_a_00317. URL https://aclanthology.org/2020.tacl-1.30/.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269/.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. MERA: A comprehensive LLM evaluation in Russian. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9920–9948, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.534. URL https://aclanthology.org/2024.acl-long.534/.

Google DeepMind. Google gemini ai update - december 2024. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/, 2024. Accessed: 2025-04-11.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL https://aclanthology.org/2022.tacl-1.30/.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.438. URL https://aclanthology.org/2020.emnlp-main.438/.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *CoRR*, abs/2410.15553, 2024. doi: 10.48550/ARXIV.2410.15553. URL https://doi.org/10.48550/arXiv.2410.15553.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174, 2023.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 5943–5959, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.479. URL https://aclanthology.org/2020.emnlp-main.479/.

Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *CoRR*, abs/2504.02768, 2025. URL https://doi.org/10.48550/arXiv.2504.02768.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.284. URL https://aclanthology.org/2021.eacl-main.284/.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12359–12374, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.760. URL https://aclanthology.org/2023.emnlp-main.760/.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.653. URL https://aclanthology.org/2020.acl-main.653/.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: Measuring massive multitask language understanding in Chinese. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.671. URL https://aclanthology.org/2024.findings-acl.671/.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.616. URL https://aclanthology.org/2022.emnlp-main.616/.

Lovish Madaan, Aaditya K Singh, Rylan Schaeffer, Andrew Poulton, Sanmi Koyejo, Pontus Stenetorp, Sharan Narang, and Dieuwke Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024.

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. Crosslingual generalization through multitask finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL https://aclanthology.org/2023.acl-long.891/.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupkes. Separating form and meaning: Using self-consistency to quantify task understanding across multiple senses. In Sebastian Gehrmann, Alex Wang, João Sedoc, Elizabeth Clark, Kaustubh Dhole, Khyathi Raghavi Chandu, Enrico Santus, and Hooman Sedghamiz, editors, *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 258–276, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.gem-1.22/.

Xenia Ohmer, Elia Bruni, and Dieuwke Hupke. From form(s) to meaning: Probing the semantic depths of language models using multisense consistency. *Computational Linguistics*, 50(4):1507–1556, 12 2024. ISSN 0891-2017. doi: 10.1162/coli_a_00529. URL https://doi.org/10.1162/coli_a_00529.

OpenAI. Mmmlu dataset. https://huggingface.co/datasets/openai/MMMLU, 2025. Accessed: 2025-04-11.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

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

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.185. URL https://aclanthology.org/2020.emnlp-main.185/.

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

Jirui Qi, Raquel Fernández, and Arianna Bisazza. Cross-lingual consistency of factual knowledge in multilingual language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.658. URL https://aclanthology.org/2023.emnlp-main.658/.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia Soltani Moakhar, Ran Tamir, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. INCLUDE: evaluating multilingual language understanding with regional knowledge, 2024. URL https://doi.org/10.48550/arXiv.2411.19799.

Eduardo Sánchez, Belen Alastruey, Christophe Ropers, Pontus Stenetorp, Mikel Artetxe, and Marta R. Costa-jussà. Linguini: A benchmark for language-agnostic linguistic reasoning. *CoRR*, abs/2409.12126, 2024. doi: 10.48550/ARXIV.2409.12126. URL https://doi.org/10.48550/arXiv.2409.12126.

Priyanka Sen, Alham Fikri Aji, and Amir Saffari. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1604–1619, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.138/.

Sheikh Shafayat, H Hasan, Minhajur Mahim, Rifki Putri, James Thorne, and Alice Oh. BEnQA: A question answering benchmark for Bengali and English. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1158–1177, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.68. URL https://aclanthology.org/2024.findings-acl.68/.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=fR3wGCk-IXp.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, André F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global MMLU: understanding and addressing cultural and linguistic biases in multilingual evaluation. *CoRR*, abs/2412.03304, 2024. doi: 10.48550/ARXIV.2412.03304. URL https://doi.org/10.48550/arXiv.2412.03304.

Mistral AI team. Cheaper, better, faster, stronger, 2024. URL https://mistral.ai/news/mixtral-8x22b. Accessed: 4-Apr-2025.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *CoRR*, abs/2406.12624, 2024. URL https://doi.org/10.48550/arXiv.2406.12624.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Douglas Teodoro, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *CoRR*, abs/2503.10497, 2025. URL https://doi.org/10.48550/arXiv.2503.10497.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/117c5c8622b0d539f74f6d1fb082a2e9-Abstract-Datasets_and_Benchmarks.html.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1131. URL https://aclanthology.org/N19-1131/.

Average EM	The first main metric we use to quantify performance for MultiLoKo is the average Exact Match score across languages, which expresses how many of the answers match one of the gold standard answers verbatim (after post-processing the answers).
Gap	The second main metric is the gap between a model's best and worst performing language. We gap to quantify the extent to which a model has achieved *parity* across languages. Because a small gap can be achieved both through parity on high scores as parity on low scores, it is most informative in combination with average benchmark performance.
Mother tongue effect (MTE)	MTE expresses the impact of asking questions in a language in which the requested information is locally salient, compared to asking it in English. A positive MTE indicates information is more readily available in the language it was (likely) present in the training data, whereas a negative mother tongue effect indicates the information is more easily accessible in English.
Locality effect (LE)	LE quantifies the effect of using locally sourced vs translated data. It is measured by computing the difference between scores for locally sourced data and translated English-sourced data. A positive LE implies that using translated English data *underestimates* performance on a language, a negative LE that using translated English data *overestimates* performance.

Table 2: **MultiLoKo metric cheatsheet.** We use several metrics to quantify model performance using MultiLoKo. This table provides a cheatsheet for their meaning.

A Related work

In this paper, we introduce a new multilingual benchmark for LLMs, that we believe addresses gaps and pitfalls in existing benchmarks. We (concisely) outlined those gaps and pitfalls and mentioned several other works related to ours in the introduction of those paper. Here, we discuss multilingual evaluation of LLMs in more detail. Specifically, we discuss what datasets recent LLM releases have used for multilingual evaluation (Appendix A.1) and what other datasets and approaches they could have used but did not (Appendix A.2).

A.1 Multilingual evaluation of LLMs in practice

While multilinguality is something frequently mentioned in the release papers or posts of recent LLM releases, the datasets for which they actual report scores is in most cases quite limited. Of the models that we evaluated for this paper, Gemini 2.0 Flash reported no multilingual scores at all; GPT4-o and Mixtral 8x22B report scores only on internally translated but not publicly available English benchmarks; Claude 3.5 Sonnet reports scores for only one benchmark – MGSM. MGSM is also the only publicly available benchmark for which Llama 3.1 reports scores, along with – also – an internally translated version of MMLU that is not publicly available. The only model that extensively reports multilingual benchmark values, on more than 10 benchmarks, is Qwen2.5 72B. We provide an overview of the multilingual benchmarks for which scores are reported for these models in Table 3.

Claude 3.5 Sonnet	MGSM (Shi et al., 2023)
Gemini 2.0 Flash	Mentions multilingual audio, no multilingual benchmarks scores reported.
GPT4-o	ARC-Easy and TruthfulQA translated into five African languages (internal benchmark), Uhura-Eval (internal benchmark).
Llama 3.1	MGSM (Shi et al., 2023), Multilingual MMLU (internal benchmark)
Mixtral 8x22B	translated ARC-C, HellaSwag and MMLU (internal benchmarks)
Qwen2.5 72B	M3Exam (Zhang et al., 2023), IndoMMLU (Koto et al., 2023), ruMMLU (Fenogenova et al., 2024), translated MMLU (Chen et al., 2023), Belebele (Bandarkar et al., 2024), XCOPA (Ponti et al., 2020), XWinograd (Muennighoff et al., 2023), XStoryClose (Lin et al., 2022), PAWS-X (Zhang et al., 2019), MGSM (Shi et al., 2023), Flores-101 (Goyal et al., 2022)

Table 3: **Multilingual evaluation of recent LLM releases, overview.** We provide an overview table of the benchmark for which scores are reported in the release papers or notes of the LLMs we evaluated in this paper. Models are sorted alphabetically.

A.2 Multilingual evaluation options for LLMs

While, as we discuss below, there are gaps and challenges with multilingual evaluation for LLMs, there are in fact many more options than is suggested by what is reported in recent releases. Below, we discuss other options for multilingual LLM evaluation.

Translated English benchmarks As mentioned earlier on, benchmarks used for LLM evaluation are often translated English benchmarks. In some cases, the benchmarks were designed to evaluate only English and translated later, such as translated MMLU (e.g. Li et al., 2024; Chen et al., 2023; OpenAI, 2025; Singh et al., 2024) or MMLU-ProX (Xuan et al., 2025), MGSM (Shi et al., 2023) or MLAMA (Kassner et al., 2021). In other cases, the benchmark was multilingual at the time of its creation, but means of creation of the non-English data was through translating English sourced data, such as Belebele Bandarkar et al. (2024), Mintaka (Sen et al., 2022), or X-FACTR (Jiang et al., 2020). Taken together, translated benchmarks span quite a range of tasks, such as question answering (Artetxe et al., 2020; Lewis et al., 2020; Qi et al., 2023; Ohmer et al., 2023), natural language inference (Conneau et al., 2018), paraphrase detection (Zhang et al., 2019), general linguistic competence (Jumelet et al., 2025), reading comprehension (Artetxe et al., 2020; Bandarkar et al., 2024) and commonsense reasoning (Ponti et al., 2020), and even instruction following (He et al., 2024). With the exception of question answering and of course instruction following, however, many of these tasks have gone (somewhat) out of fashion for LLM evaluation, a trend which is mirrored also in the usage of their multilingual counterparts. As mentioned before, translated benchmarks have the advantage of containing parallel data, allowing for some form of comparability across languages, but are English-centric in content and may suffer from translationese (see e.g. Romanou et al., 2024; Chen et al., 2024, for a recent discussion of this).

Multilingual benchmarks sourced from scratch Though much rarer, there are also benchmarks that are created independently for each language they include. Clark et al. (2020) release a question answering dataset separately sourced for 11 different languages, with a protocol relatively similar to ours. In a different category, Hardalov et al. (2020), Zhang et al. (2023) and Romanou et al. (2024) and Sánchez et al. (2024) do not *create* benchmark data, but instead collect existing exam or competition questions from official human exams. In case of Zhang et al. (2023), the exams are graduation exams of primary, middle and high school; Hardalov et al. (2020) includes official state exams taken by graduating high school students, which may contain parallel pairs in case countries allow examinations to be taken in multiple languages; Romanou et al. (2024), cover academic exams at middle and high school and university level, professional certifications and licenses, and exams to obtain regional licenses. Sánchez et al. (2024) instead focus on questions from the International Linguistic Olympiad corpus. Lastly, as part of their study Ohmer et al. (2023) create a dataset called SIMPLE FACTS, containing factual questions created through a shared template filled in with language specific factual data.

Consistency evaluation A rather different approach to assess multilinguality in LLMs is to focus not on accuracy across different languages, but to consider whether predictions are *consistent* across languages. This tests knowledge and skill transfer between languages more explicitly. Two recent examples of studies incorporating consistency-based evaluations on factual knowledge questions are Qi et al. (2023) and Ohmer et al. (2023). Qi et al. (2023) focusses specifically on sample-level consistency of answers across different languages, requiring existing parallel benchmarks. Ohmer et al. (2023), instead, ask models to translate benchmark questions themselves before answering them again. This can, with some caveats, be applied to any existing monolingual benchmark, but – requiring multiple steps – it is more involved an a paradigm, and is somewhat bottlenecked by the translation ability of the model to be evaluated.

Translation as a proxy for multilinguality Another, more implicit method to assess multilinguality in LLMs is to evaluate their ability to translate from one language to another. This approach was famously used by Brown et al. (2020), but has not been common since.

Monolingual non-English evaluation In our discussion, we have focussed on multilingual evaluation options that cover multiple other languages. After all, a benchmark to evaluate models on Bengali (e.g. Shafayat et al., 2024) or Arabic (e.g. Alwajih et al., 2024) can contribute to multilingual evaluation when combined with other benchmarks, but does not so on its own. Because such benchmarks are usually created by language experts for the respective languages, they usually target locally relevant skills and knowledge and are likely of higher quality than benchmarks created for many languages simultaneously (either through translation or from scratch). Yet, composing a suite including many languages that allows direct comparisons between languages remains challenging. We believe such benchmarks can be important for multilingual evaluation in LLMs, but will not further discuss benchmarks focussing on individual languages or very small sets of languages within one family here.

B Additional dataset statistics

For reference, we provide a few dataset statistics beyond the main results in the paper.

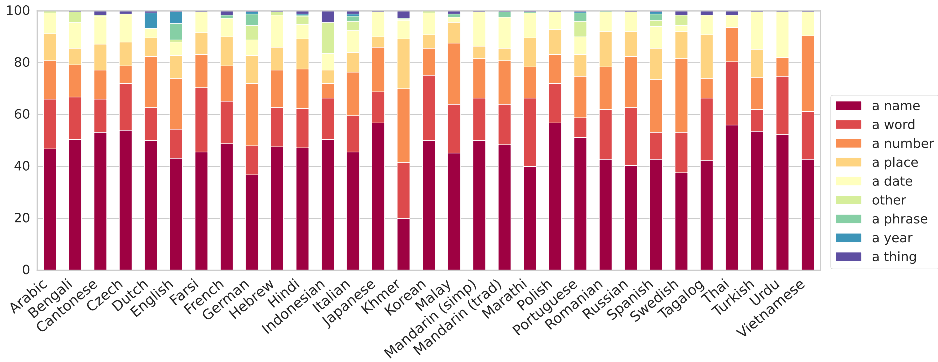


Figure 6: **Distribution of output types on the dev split.** We show the normalised distribution of correct output types across languages, ordered (from bottom to top) by average frequency. Rare output types that occur only a few times are mapped to the category 'other'.

Output type distribution In Figure 6, we show the per-language distribution of output types for MultiLoKo dev split.[6] We mapped very rare output types, such as 'a quantity', 'a period of time' or 'letter' to 'other', for plotting purposes. We can see that *name* is the most common output type across languages, followed by the generic output type *a word* and *number*. Also *place* and *date* are relatively common output types, whereas most other output types occur very infrequently or only for a handful of languages.

Input and output length In addition to that, we show the average question – and output lengths of human-translated the locally sourced questions to English in Figure 7. While there is some variation in particular in question length, the lengths of the answers are relatively consistent. The average answer length is around 2, combining one-word answers with (usually) longer names.

C Model runtime and compute resource information

In this section we provide details about the runtime for non-API models (LLaMa 3 family of models, Qwen and Mixtral).

Hardware	All models were run on one node of 8xH100 80GBs, 1 TB of RAM, with the exception of LLama 3.1 405B, for which we used 2 nodes.
Precision	All models were run on **bf16** precision.
Runtime	Excluding model loading time, all models took <10 minutes to complete a partition of the dataset. A partition is defined as 250 examples, and could be either dev, test, or human/machine translated version of those. Overall, < 1 hour is required to run every single example in our dataset through an LLM with this setup.

For iterating over prompts and debugging, we used a reduced dataset of 50 examples, and a 70B model, which overall took less than 1 hour of total compute resources.

[6]Because the test split is blind, we do not report the distribution of output types here.

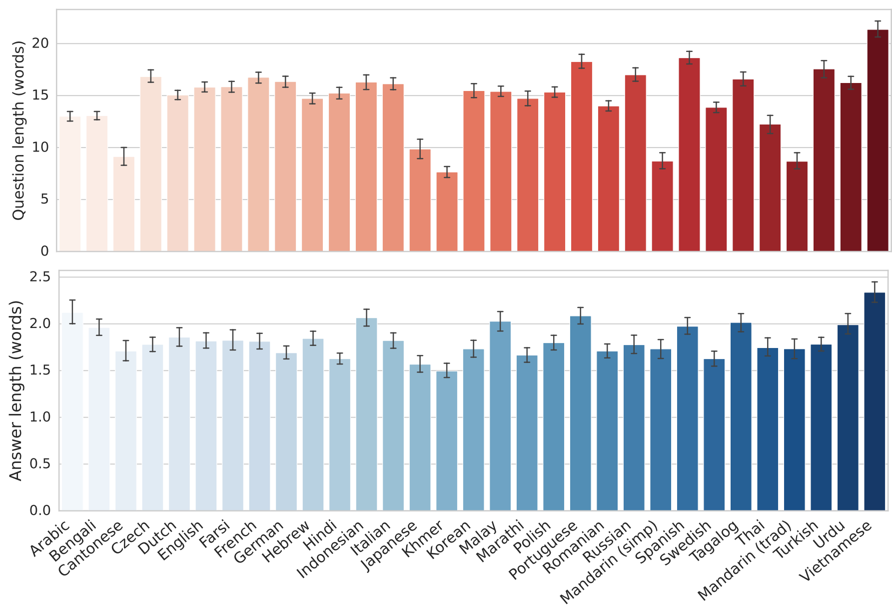


Figure 7: **Average question and answer lengths.** We show the per-question average length (in words) of the locally-sourced questions and answers, human-translated into English.

D Differences between languages

So far, with the exception of MTE and parity scores, we have primarily looked at results averaged across languages. Now, we consider language-specific results in a bit more detail.

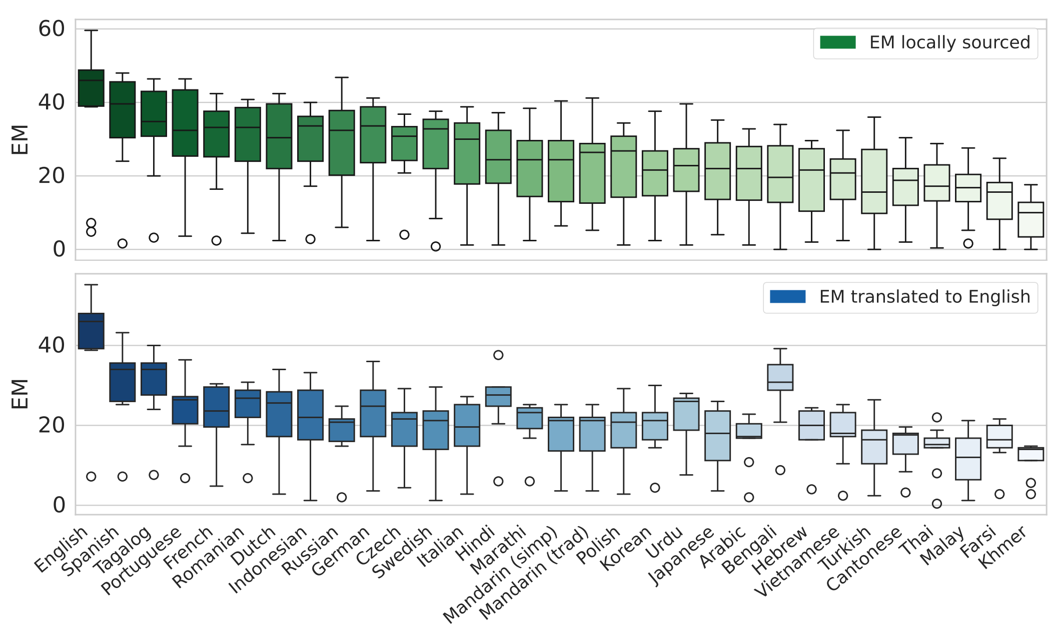


Figure 8: **Average EM per language dev, in mother tongue and English.** Top: Average EM on locally sourced data. Bottom: Average EM on locally sourced data, translated to English.

D.1 Language difficulty on locally sourced data

First, in Figure 8 (top), we show average results for all languages on all locally sourced data. In broad strokes, the order of difficulty is correlated with how low- or high- resource a language is considered: while languages such as French, English and Spanish occur at the easier end of the spectrum, we find Farsi, Khmer and Malay among the most difficult languages. There are a few notable exceptions: on average the second highest scoring language in our benchmark is Tagalog. While it is difficult to judge why without doing a detailed analysis on the questions, we hypothesise that the questions asked by the Tagalog language experts are simply less complex than the questions of other languages.

D.2 Separating language difficulty from language proficiency

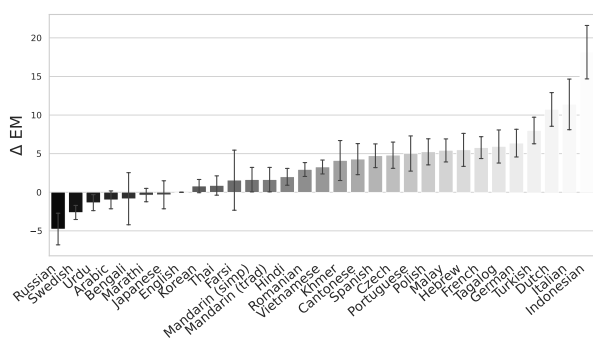
In an attempt to distinguish *data difficulty* from *language proficiency*, we consider also the difficulty of the locally sourced data translated to English. While this conflates data difficulty and transfer (see § 5.2), it still gives us some indication of the extent to which low performance in languages is caused by poor language proficiency versus data difficulty. In the bottom half of Figure 8, we show the model performances as computed on the locally sourced data *translated to English*. The correlation between these two language difficulty rankings between these setups is 0.79. When comparing the ranks of the various languages, only a handful of languages shift more than a few places. Specifically, Bengali (26->4), Urdu (26->12), and Hindi (14->5) all decrease substantially in difficulty rank. The fact that they are comparatively easier in English suggests that for those languages proficiency may be a larger problem than data difficulty. On the other hand, only Russian (7->21) shows a drop of more than 5 places.

E Human versus machine translation

In this section, we consider the impact of using machine- or human-authored translations. To do so, we look at the differences in EM scores between machine and human translated data for the various languages, taking the human translations as the 'gold standard' (i.e. we consider human translated EM - machine translated EM). We show the results in Figure 9.

Model	R	min Δ	max Δ	avg Δ
Gemini 2.0 Flash	0.80	-10.00	21.60	4.35
Llama 3.1 405B	0.83	-4.40	18.80	5.82
GPT4-o	0.85	-6.00	21.60	4.46
Llama 3.1 405B Chat	0.80	-10.40	22.40	3.08
Llama 3.1 70B	0.77	-7.60	22.00	4.59
Claude 3.5 Sonnet	0.90	-9.60	20.80	2.84
Llama 3.1 70B Chat	0.87	-6.00	20.00	3.12
Mixtral 8x22B	0.91	-3.20	20.00	4.13
Qwen2.5 72B	0.83	-4.00	16.80	3.47
Mixtral 8x22B-it	0.92	-4.80	12.40	2.41
Qwen2.5 72B instruct	0.80	-0.80	3.20	0.36

(a) Language difficulty stats across human- and machine translations



(b) MT vs human translations

Figure 9: **Machine versus human translations dev.** Left: Per-model rank correlation between language difficulty between MT and human translations, and min, max and average difference between the two conditions. Right: Difference between EM computed on human- and machine-translated data (human score - machine score), per language.

In Figure 9a we show the rank correlations of the difficulties of the various languages per model, as well as the min, max and average drop from human to machine translations. We see the that, at the model level, using machine translations rather than human translations results in a systematic undervaluation of the model scores: there is not a single model for which the 'drop' from human to machine translations is negative on average. In part, this is may be a result of the previously observed lack of knowledge transfer effect. That the drop is not substantially lower for models with better transfer, however, suggests that the more impactful factor is the quality of the machine translations, that may at times result in unanswerable questions.

In terms of model rankings, the difference between machine and human translations is minor: the model rankings between the two conditions have a rank correlation of 0.97 on the dev split, with only three local swaps (2&3 and 5&6 and 8&9) of models that did not have statistically different scores to begin with. This suggests that to compare models, using machine translation can be an acceptable alternative to human translations, as the mis-estimation is systematic across models.

Considering the effect across languages, we observe that even though the average drop is positive, for virtually all models there are at least some languages for which performance increases when MT is used, in some cases with even more than 10 points. For a handful of languages – specifically Russian, Swedish and Urdu – this is also true across models (see Figure 9b). While the overall rank correlation is high for language difficulty (0.88), it thus still urges caution in using machine translated data for language improvement prioritisation.

F Limitations

In this last section, we discuss various limitations of our work.

Local relevance In our sourcing protocol, we explicitly sought to create questions locally relevant to the respective languages. It is important to notice, however, that some languages, such as English, Spanish, Portuguese, Chinese, French and to a lesser extent German and Dutch cover a wide variety of cultures. We did not separately control for that and the data for those languages thus likely comprises a mix of different locales.

Data quality Building a bias-free evaluation datasets with few mistakes is not an easy feat. Even though we implemented several rounds of quality checks in our data collection pipeline, when looking at outputs we still incidentally found mistakes in the data or answers. We fixed some of these mistakes as we encountered them, but it is quite likely that more such mistakes occur in the dataset. It is also important to point out that we are less likely to spot such issues for languages that we do not understand at all, potentially creating a bias towards the set of languages for which we have a rudimentary understanding. Overall, however, we believe that the pipeline we designed assures a dataset of high quality. Of course, we welcome reports of mistakes spotted by others in the data.

Evaluation Because MultiLoKo is a generative benchmark, computing scores requires comparisons of a generated answer with a set of gold answers. A first obstacle to this method of evaluation is that it is hard to create an exhaustive list of correct short-form answers. This is especially true when the correct answer is not a number, date, title or something else that can be expressed only in a few ways. In addition to that, it is hard to incentivise LLMs to produce concise answers. Even when instructed to answer with only a number / date / name / title, they may respond with a full sentence, add a reasoning trail to their answer, or add words beyond the minimal answer in a different fashion. We addressed such issues that were systematic in post-processing (see Appendix G), but it is hard to a priori catch all the ways that LLMs may deviate from the requested protocols. In some cases, we found additional post-processing steps that increased the scores of some models only later in the process, because scores for particular languages looked suspiciously low. For instance, we had not initially realised that our punctuation stripper did not strip punctuation in Urdu, which specifically influenced GPT4-o and Gemini. We considered several other metrics as well as judges, but eventually found that EM provided the clearest and least biased signal. It remains, however, a challenge to evaluate chatty LLMs completely independently from their ability to follow instructions.

Wikipedia as information source MultiLoKo, as several other both multilingual as well as monolingual benchmarks, uses Wikipedia as main source of information. This has the advantage that Wikipedia has a large coverage across many different languages and the information is considered to be of high quality. It also facilitates comparable sourcing across languages. Of course, it also poses limitations. For one, it still provides a bias to the specific topics that can be included, that are usually primarily knowledge based. In fact, MultiLoKo is indeed a knowledge benchmark; it does not consider other types of skills. Secondly, and perhaps more importantly, Wikipedia is a a corpus frequently used in the training data of LLMs. The fact that MultiLoKo is a challenging benchmark even given that (multilingual) wikipedia is likely included in the training data of most of the LLMs evaluated suggests that this is not a large issue at the moment. However, it is very possible that MultiLoKo can be 'hacked' relatively easily simply by strongly oversampling multilingual wikipedia data.

G Instruction following

To facilitate evaluation, we instruct models to answer question with only a number/place/etc. Overall, we found that base models (with a five-shot template) are much better at abiding by this instruction than chat models, which exhibit a number of pathologies. While some of those can be caught with appropriate post-processing (see Appendix H, this is not the case for all issues. Below, we provide a summary of the main instruction-following issues we encountered with chat models.

False refusals Sometimes chat models refuse to provide an answer when the question is falsely perceived to be inappropriate (e.g. when the question asks about someone aged younger than 18).

Producing full sentences Another issue we observed is that chat models would provide a full sentence answer, rather than a single word or phrase (e.g. *Which year was Francisco Franco born? Produce a year only. – Francisco Franco was born in 1936*). Such full-sentence answers make exact match rating impossible. The effect is not consistent across languages and happens only for some of the examples, without any discernable pattern, and therefore difficult to address completely with post-processing.[7]

Spurious addition of "answer is" Likely due to overtraining on MMLU style tasks, Models such as OpenAI's GPT4 and Gemini 2.0 preface the vast majority of the answers in English with "answer is" or "X answer is X" where X is the desired correct response. This is remarkable, because it is essentially a repetition of the end of the prompt. However, it is easy to fix in post-processing.

Japanese specific issues In Japanese, in general it is not polite to answer with incomplete sentences. As such chat models often append the copula verb "desu" to the answer, making exact match unsuccessful. We are able to fix this in postprocessing.

Claude 3.5 Sonnet issues We were unable to make Claude 3.5 Sonnet follow the instructions to produce just an answer in English. It seemed to engage in a long chain-of-thought reasoning style response which we were unable to reliably parse. This issue only manifests in English and only with Claude. For this reason, we exclude Claude 3.5 Sonnet from our knowledge transfer results, as it would make the average lack of knowledge transfer from non-English languages to English more severe than they are.

H Post-processing details

We perform the following post-processing for both the reference answers and the answers produced by the model:

- Remove leading and trailing whitespaces.
- Remove punctuation.
- Lowercase everything.

We perform the following additional post-processing for pretrained models:

- Remove leading "Answer:" or "A:" or the non-English equivalent from the output.
- Remove everything after the first newline.

We perform the following additional post-processing for postrained models:

- Remove leading "answer is:"
- Detect the pattern "X answer is X", where X is the desired answer, and strip the unnecessary part in the middle.
- Remove training "desu" in Japanese.

[7] Using a judge-LLM may to some extent address this problem, but at the expense of other issues (e.g. Thakur et al., 2024).

I Annotation instructions

Our annotation pipeline contains five stages: 1) locality rating, 2) question generation 3) question review, 4) question answering, and 5) translation. Below, we provide the annotation instructions for each of these stages.

I.1 Locality rating

To narrow-down the initial selection of paragraphs – sampled from the top-rated Wikipedia pages of the respective locales – the first step in our annotation pipeline is *locality rating*. Given a paragraph, we ask annotators to rate whether the paragraph is locally relevant to the particular locale, on a likertscale from 1 to 5, where 1 refers to extremely local and relatively obscure topics very specifically related to the specific language or locale and with little international recognition and 5 to globally well-known topics. We also ask annotators to disregard pages about inappropriate or politically sensitive topics. The rubric for locality annotation can be found in Table 4. We disregard everything with a locality rating of 3 or lower.

	Description	Example
1.	**Extremely local and relatively obscure.** Content that is of interest only to a small, localized group, such as a specific town, region, or community. These topics are typically obscure and not widely known beyond their immediate area.	Local radio stations, small town historical events, regional businesses, or niche local cultural practices.
2.	**Regional interest.** Topics that have some relevance beyond a specific locality but are still primarily of interest within a particular region or country.	State or provincial politicians, regional cuisine, local sports teams, or medium-sized companies with regional influence.
3.	**National Significance.** Content that is widely recognized within a single country, but relatively unknown internationally.	National politicians (not internationally known), popular national media figures, major corporations within a country, or significant national historical events.
4.	**International recognition.** Topics that are recognized and have relevance in multiple countries but may not be universally known across the globe. These topics often have international influence and are likely to be covered in international media, though their impact may vary by region.	International brands which may be recognized in more than one country, celebrities with some international reach, significant cultural movements, or political conflicts with some awareness on the international stage.
5.	**Global prominence.** Content that is widely recognized and relevant across a large number of countries around the world. These topics have a global impact or appeal and are likely to be well-represented in media across diverse cultures and regions.	Globally famous celebrities (e.g., Cristiano Ronaldo), multinational corporations (e.g., Apple), major world events, or universally recognized cultural icons.

Table 4: **Rubric for locality rating task.** In the locality rating task, we ask the annotators to rate paragraphs with respect to how locally relevant the topic is to the locale.

I.2 Question generation

The second and main annotation step in our pipeline is the step in which we ask annotators to generate questions about sampled paragraphs. We ask annotators to generate a challenging question with a short answer. The answer should be easy to evaluate with string-matching metrics, the questions should not be open-ended or have many possible correct answers, be ambiguous or subjective, and the expected short answer should be concise. To ensure difficulty, we ask that answering the question requires combining information from different parts in the accompanying text; It should not be answerable by mere regurgitation of a single sentence. We furthermore ask that the question is formulated such that its answer will not change over time (e.g. not 'How many medals has Sifan Hassan won', but 'How many medals has Sifan Hassan won between 2018 and 2022 (including)'), and that the question is answerable also without the article (e.g. not 'How many tv shows did the person in this article produce?'). To facilitate validation checks in the next round, we also ask that the question authors write a longer answer to explain how they arrived at the short answer. We also ask the question authors to annotate what is the type of the correct answer (e.g. number, name, date, etc) In the pilot, we observed that – for some languages – the vast majority of questions were questions that required some form of numerical reasoning. Because the intention of the benchmark is to address knowledge more than reasoning, we afterwards restricted the number of numerical questions to 10%. Similarly, we asked question authors to avoid yes/no questions.

I.3 Question review

In the first round of question review, we asked annotators from a different provider to judge whether the questions abide by the rules provided to the question authors. All question reviewers are native speakers. Specifically, we ask them to check if:

- The question pertains to a locally relevant topic
- The question is clear and undestandable, and not subjective
- The question has a clear and concise answer
- If there are multiple possible variations of the answer possible (e.g. 'Dick Schoof' / 'Minister Dick Schoof' / 'Prime Minister Dick Schoof' / etc), all versions of the answer are provided.
- The question and answer are in the correct language
- The question is understandable without the article
- That the answer to the question will not likely change in the near future

When a question can be fixed with a minor change (e.g. add a time indication to make sure an answer will not change in the near future, or add an extra answer version), we ask the question reviewers to implement this fix and describe it. In the pilot round, we use the annotator feedback to finetune our annotation protocol and provide feedback to the question-authors. During the rest of the data collection, we simply disregard questions that are not useable as is or can be corrected with minor changes.

I.4 Validation through question answering

In the last stage of our question generation pipeline, we have additional annotators answer the sourced and reviewed question. The goal of this validation task is to confirm that the questions are answerable, correct, non-ambiguous when read by individuals other than the original question author, and that all possible versions of the answers are included. For each question, we ask two additional annotators to first answer the question, using the snippets the questions were sourced from for context. After they have answered the question, they are shown the list of reference answers written by the original author of the question as well as the rational they provided, and we ask them to reflect upon the answer they gave themselves. If their answer did not match any answer in the original reference list, we ask them to either add their answer to the list if it is semantically equivalent to their own answer or indicate which answer they believe to be correct, their own or the original answer. We disregard all questions where at least one annotator disagrees with the original question author.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction we outline the problem (lack in adequacy of datasets to assess multilinguality in LLM), describe how we solve it (propose a new dataset) and give an outline of our main results using this dataset.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: It does, both throughout the work and in more detail in Appendix F.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

   Justification: There are no theoretical results in this paper

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide details about the parameters we ran the experiments with and release all prompt templates and post-processing code. Note though that API-based results are non-deterministic even with generation temperature of 0, so exact reproduction of our results will likely not be possible.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, we release the data and code for reproduction at `https://github.com/facebookresearch/multiloko`

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all details regarding our decision to split the data in the main paper. We did not train any models.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Where appropriate, we indicate confidence intervals and significance values (typically 2 times standard error) and describe what they are.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We outline the requested information in Appendix C.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We believe that the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We briefly discuss the societal importance of multilingual evaluation in the introduction, but not much beyond that.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: We are not releasing any models and believe that the data popses no risk for misuse, beyond training models directly on the data which would render the benchmark itself useless. We have furthermore password protected the data to prevent accidental scraping, and have included a license asking others not to mirror it.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We provided references for all data and models used and abided by their license terms.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We have elaborately documented both data collection processes and the data itself.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We used an external vendor for data collection but did no crowd-sourcing experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our study did not have participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.