When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages

Tyler A. Chang[a,c] **Catherine Arnett**[b] **Zhuowen Tu**[a] **Benjamin K. Bergen**[a]
[a]Department of Cognitive Science
[b]Department of Linguistics
[c]Halıcıoğlu Data Science Institute
University of California San Diego
{tachang,ccarnett,ztu,bkbergen}@ucsd.edu

Abstract

Multilingual language models are widely used to extend NLP systems to low-resource languages. However, concrete evidence for the effects of multilinguality on language modeling performance in individual languages remains scarce. Here, we pre-train over 10,000 monolingual and multilingual language models for over 250 languages, including multiple language families that are under-studied in NLP. We assess how language modeling performance in each language varies as a function of (1) monolingual dataset size, (2) added multilingual dataset size, (3) linguistic similarity of the added languages, and (4) model size (up to 45M parameters). We find that in moderation, adding multilingual data improves low-resource language modeling performance, similar to increasing low-resource dataset sizes by up to 33%. Improvements depend on the syntactic similarity of the added multilingual data, with marginal additional effects of vocabulary overlap. However, high-resource languages consistently perform worse in multilingual pre-training scenarios. As dataset sizes increase, adding multilingual data begins to hurt performance for both low-resource and high-resource languages, likely due to limited model capacity (the "curse of multilinguality"). These results suggest that massively multilingual pre-training may not be optimal for any languages involved, but that more targeted models can significantly improve performance.

1 Introduction

Multilingual language models have been a fixture of natural language processing (NLP) research nearly since the introduction of Transformer language models (Devlin et al., 2019; Conneau et al., 2020a). These models are often pre-trained on over 100 languages simultaneously, and they are widely used for NLP tasks in low-resource languages (Adelani et al., 2021; Ebrahimi et al., 2022; Hangya et al., 2022; Imani et al., 2023), cross-lingual transfer learning (Pires et al., 2019; Conneau et al., 2020a), and multilingual text generation (Lin et al., 2022; Scao et al., 2022). However, while multilingual language models produce strong results across many languages, multilingual pre-training work almost exclusively focuses on pre-training a small number of models with some fixed distribution over languages (e.g. mBERT, XLM-R, XGLM, and BLOOM; Devlin et al., 2019; Conneau et al., 2020a; Blevins et al., 2022; Lin et al., 2022; Scao et al., 2022). This distribution over languages typically favors high-resource languages spoken in regions with high economic influence (Bender, 2011; Joshi et al., 2020).

Thus, it is largely unknown how different pre-training language distributions, such as different quantities of multilingual data or different selections of languages, affect multilingual language model performance in different languages. Multilingual models have been studied extensively during inference and fine-tuning (Pires et al., 2019; Conneau et al., 2020b; Karthikeyan et al., 2020; Winata et al., 2021; Chai et al., 2022; Alabi et al., 2022; Guarasci et al., 2022; Winata et al., 2022; Wu et al., 2022; Eronen et al., 2023), but these studies generally rely on the same sets of pre-trained models. For pre-training, there is mixed evidence for the benefits of multilingual vs. monolingual data (Conneau et al., 2020a; Wu and Dredze, 2020; Pyysalo et al., 2021; §2). As multilingual language models are increasingly used without task-specific fine-tuning (e.g. for text generation; Scao et al., 2022; Lin et al., 2022),[1] it is critical to understand how multilingual pre-training affects raw language modeling performance in individual languages.

In our work, we investigate the effects of different multilingual pre-training distributions on

[1]The multilingual text generation capabilities of recent commercial models also indicate likely multilingual pre-training (OpenAI, 2023; Google DeepMind, 2023).

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4074–4096
November 12-16, 2024 ©2024 Association for Computational Linguistics

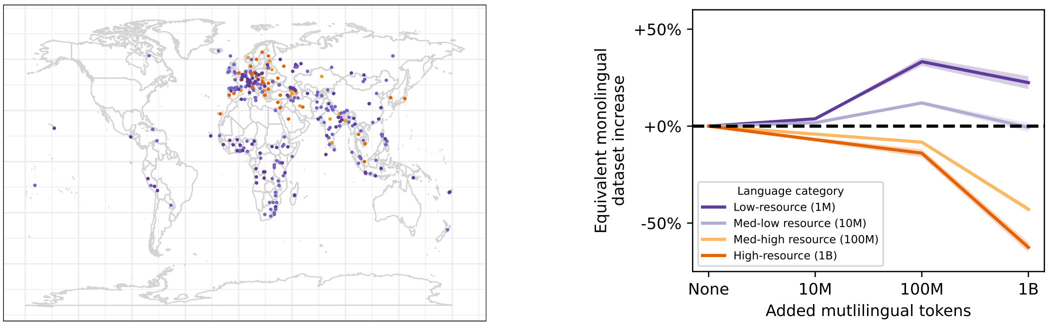

Figure 1: Left: Map of the 252 languages used in our study. Right: Effects of adding multilingual pre-training data in similar languages, for low-resource (1M token) through high-resource (1B token) languages in small models. Effects are quantified using the estimated monolingual dataset size that would achieve similar performance. Adding 1B tokens of multilingual data is similar to adding 22% (low-resource) or removing 63% (high-resource) of the monolingual dataset. Shaded regions are 99% confidence intervals for the mean across languages.

language modeling performance in 252 languages. Our main contributions are:[2]

- We pre-train over 1900 monolingual baseline models for 252 languages, and we estimate loss in each language based on monolingual dataset size (§4). We use these estimates to quantify multilingual model performance in individual languages (§4.3).
- We pre-train over 8400 multilingual language models, and we evaluate how performance in individual languages varies as a function of monolingual dataset size, multilingual dataset size, linguistic similarity of the training languages, and model size (up to 45M parameters; §5).
- We find that moderate amounts of multilingual data improve performance for low-resource languages, similar to increasing low-resource dataset sizes by up to 33% (§6.1). These improvements depend primarily on the syntactic similarity of the added multilingual data, with marginal additional effects of lexical (vocabulary) similarity.
- We find that multilingual data consistently hurts high-resource language performance, similar to reducing dataset sizes by over 85% in some cases (§6.2). Likely due to limited model capacity, as dataset sizes increase, adding multilingual data begins to hurt performance for both low-resource and high-resource languages (the *curse of multilinguality*; §2).

Thus, the benefits of multilinguality on raw language modeling performance appear restricted to cases where both (1) the model targets performance in low-resource languages and (2) the model has enough capacity for the added multilingual data. If these assumptions hold, the multilingual data should be from languages that are linguistically similar to the target low-resource languages. However, when optimizing performance for high-resource languages, multilingual models are likely to degrade performance in individual languages.

2 Related Work

The curse of multilinguality. To extend language models to low-resource languages, researchers often train a single model on a large number of languages, including low-resource languages (Devlin et al., 2019; Conneau et al., 2020a; Lin et al., 2022; Scao et al., 2022; Imani et al., 2023). Oftentimes, better performance is observed when languages are either closely related or focused in a specific region (Kakwani et al., 2020; Ogueji et al., 2021; Ogunremi et al., 2023). However, Conneau et al. (2020a) find that pre-training on an excessive number of languages hurts model performance in each language, evaluating five subsets of languages on downstream tasks in 16 languages. This phenomenon is known as the *curse of multilinguality* or *negative interference* (Wang et al., 2020). Indeed, monolingual language models often have better language modeling performance than massively multilingual models (Pyysalo et al., 2021). However, Rust et al. (2021) find that the curse of multilinguality may simply be a result of lower quality tokenization per language. Further contradicting the curse of multilinguality, Wu and Dredze (2020) find that for low-resource languages, multilingual pre-training does improve downstream task perfor-

[2]Code is available at: `https://github.com/tylerachang/curse-of-multilinguality`

mance relative to monolingual pre-training, and Fujinuma et al. (2022) observe better cross-lingual transfer performance when a wider variety of languages is seen during during pre-training. Thus, the precise effects of multilinguality on low-resource and high-resource languages remains unclear.

To isolate these effects, we evaluate language modeling performance in 252 languages while systematically varying monolingual dataset size, multilingual dataset size, model size, and linguistic similarity of added languages during pre-training. This contrasts with previous studies that focus on individual (massively) multilingual models such as mBERT or XLM-R. Our approach allows us to determine how such models perform after varying pre-training languages and language distributions.

3 Collecting a Multilingual Dataset

Conducting controlled multilingual language modeling experiments requires a large multilingual dataset. Notably, broad language coverage is a consistent issue in NLP research (Bender, 2009, 2011; Joshi et al., 2020; Blasi et al., 2022). Here, we collect text corpora from 24 multilingual data sources such as OSCAR (Ortiz Suárez et al., 2019; Abadji et al., 2021), Wikipedia (Wikipedia, 2023), and NLLB (Costa-jussà et al., 2022). Our sources are reported in §A. We merge the corpora per language, and we deduplicate repeated sequences of 100 UTF-8 bytes (Lee et al., 2022). Restricting each language to a maximum of 1B tokens, our dataset contains 41.4B tokens in 1572 languages. This includes 252 languages with the required 1.5M tokens for our language modeling study. Despite this fairly stringent token requirement, our 252 languages cover five continents, 29 language families, and 30 scripts (i.e. writing systems). Figure 1 shows a geographic map of our 252 languages, using coordinates from Glottolog (Hammarström et al., 2023). Our list of languages with corresponding token counts is reported in §G.

4 Monolingual Baselines and Metrics

To study effects of multilinguality on language modeling performance in individual languages, we first need a method to quantify performance in those languages. Thus, we pre-train monolingual baseline models for our 252 languages, to use as comparison points for multilingual models. For each language L, we estimate the number of monolingual tokens in L required to achieve a given level of performance in L with a given model size. We later use this estimated number of monolingual tokens as an interpretable performance metric for multilingual models.

4.1 Model Architectures and Pre-Training

We pre-train autoregressive GPT-2 Transformer language models from scratch (Radford et al., 2019) with three sizes from Turc et al. (2019): tiny (4.6M parameters), mini (11.6M parameters), and small (29.5M parameters). For each language, we pre-train models with four dataset sizes when available: 1M, 10M, 100M, and 1B tokens, not including 500K tokens for evaluation in each case. We call these dataset sizes low, med-low, med-high, and high resource respectively. We have 252 languages with at least the low-resource dataset size, 167 with med-low resource, 48 with med-high resource, and 28 with high-resource. Resource categories for all 252 languages are included in §G. Hyperparameter details are reported in §C.

Monolingual tokenizers. We train a monolingual SentencePiece tokenizer with maximum vocabulary size 32K for each of our 252 languages (Kudo and Richardson, 2018), and we fix this tokenizer for all models pre-trained for that language. We train each tokenizer on 10K randomly-sampled lines of text in the language; for languages where more lines are available, the 10K-line tokenizers have reasonable vocabulary overlap with tokenizers trained on more lines (§B). For example, a 10K-line tokenizer on average covers 93.7% of the 4K most frequent tokens in the vocabulary of a 10M-line tokenizer. We restrict tokenizer training to 10K lines for all languages to control for tokenization quality across languages.

4.2 Perplexity and Log-Likelihood

As an initial performance metric, we compute the log-likelihood assigned by a language model $\mathcal{M}$ to the unseen evaluation dataset for language L. Each of our monolingual models is evaluated on its corresponding pre-training language, but these methods also apply to our multilingual models (which each have a tokenizer fixed for one target language; §5). Averaging over tokens, evaluation log-likelihood is equivalent to negative log-perplexity, mean token log-probability, or the negative of the language model's cross-entropy loss (Equation 1). Because our tokenization is fixed across all models with a given target language, log-likelihoods are comparable within each target language. Higher log-likelihood scores indicate better language modeling performance, they are predictive of model performance on other natural language tasks (Xia et al., 2023), and they can be computed even for languages without any labeled data.

Although log-likelihood scores are comparable for models with the same target language, they vary substantially across languages. This can be due to features of individual languages, their datasets, or their tokenization (Gerz et al., 2018). Thus, when model $\mathcal{M}$ is pre-trained on language L, we subtract the log-likelihood score of the baseline tiny monolingual model ($\mathcal{M}_0$) trained on 1M tokens for that language, obtaining a relative log-likelihood as follows:

$$\text{mean}_w\big(\log_2 P_{\mathcal{M}}(w)\big) - \text{mean}_w\big(\log_2 P_{\mathcal{M}_0}(w)\big) \quad (1)$$

Here, w are tokens in the evaluation dataset for L. As is standard, token probabilities are produced by the language models $\mathcal{M}$ and $\mathcal{M}_0$ based on preceding context (Brown et al., 2020). Equation 1 is then equivalent to the log-odds of observing the evaluation dataset for L using the model $\mathcal{M}$ versus $\mathcal{M}_0$. A relative log-likelihood of ℓ indicates that $\mathcal{M}$ assigns the evaluation dataset 2^ℓ times the likelihood assigned by $\mathcal{M}_0$. Equivalently, $\mathcal{M}$ has perplexity 2^ℓ times lower than $\mathcal{M}_0$. In future sections, log-likelihoods refer to relative log-likelihoods that account for the target language baseline.

4.3 Estimating Monolingual Token Counts

Relative log-likelihoods are difficult to interpret when quantifying language model performance in practice; a log-likelihood change of 1.0 does not have concrete practical implications. Log-likelihoods are also difficult to compare across model sizes (§D). Therefore, when evaluating multilingual models in later sections, we quantify performance in a language L as the estimated number of monolingual tokens in L that would achieve the same log-likelihood with the same size model. Measuring model performance in terms of estimated monolingual token counts allows us to quantify the effects of adding multilingual pre-training data across languages and model sizes.

Estimating monolingual token counts for models across 252 languages is nontrivial. Previous work has found that language modeling loss (negative log-likelihood) has a power law relationship with dataset size (Kaplan et al., 2020). Indeed, we find

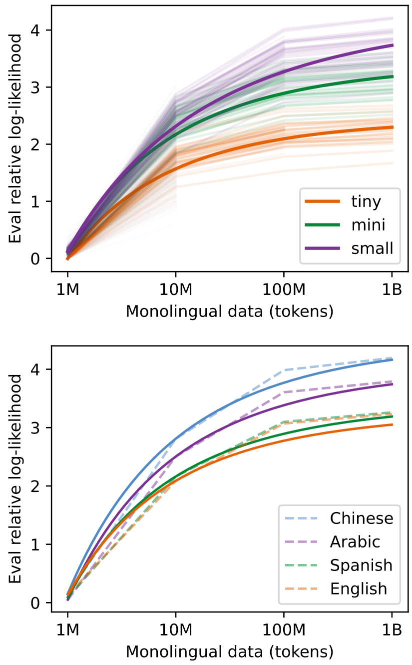

Figure 2: Curves predicting monolingual model performance from dataset size. Top: Curves fitted to all languages for each model size. Bold lines are fitted curves, and lighter lines are ground truth curves for individual languages. Bottom: Sample language-specific curves for small models, extrapolating from only two data points (1M and 10M tokens). This still produces reasonable estimates for 100M and 1B tokens. Bold lines are estimated curves, and dashed lines are ground truth values.

that $-ax^{-b} + c$ provides a good fit on average to relative log-likelihood in all 252 languages, where x is the monolingual dataset size in $\log_{10}$ tokens (Figure 2, top). However, there is significant variability in the log-likelihood vs. dataset size curve across languages. For high-resource languages, we can fit a language-specific power law to the data points for 1M, 10M, 100M, and 1B tokens. For lower-resource languages, there are too few data points to fit the power law from scratch (e.g. three power law parameters with two data points). For these languages, we fix a as the median parameter value from languages where the curve can be fit. Using this, we fit a monolingual log-likelihood vs. token count curve for each language in each model size (Figure 2, bottom; details in §D).

These curves produce reasonable estimates for the number of monolingual tokens required to achieve a given level of performance in a language L (§D). Even when token estimation accuracy is imperfect, our estimated monolingual token count is always a monotonic increasing function of eval log-likelihood, and thus performance rankings between models are preserved. In future sections, we measure the performance of a multilingual model with target language L in terms of the estimated number of monolingual pre-training tokens in L that would achieve the same performance.

5 Pre-Training Multilingual Models

Finally, we pre-train multilingual language models that vary along four dimensions: monolingual data quantity, added multilingual data quantity, model size, and linguistic similarity of the added languages. Each multilingual model is pre-trained with a specified target language, keeping monolingual tokenization for that language fixed during both pre-training and evaluation. The multilingual models are pre-trained identically to the monolingual baselines in §4, except with added multilingual data (10M, 100M, or 1B tokens). The multilingual data is randomly interspersed with the monolingual pre-training data in the target language. Target language evaluation loss curves are included in §C. In total, we pre-train 8454 multilingual language models ranging from 8M to 45M parameters.

Multilingual tokenizers. Perplexity and log-likelihood evaluations within a language L are only comparable when they use the same tokenizer. Thus, we must keep the monolingual tokenizer fixed for any model evaluated on L. However, fixing tokenization for multiple languages simultaneously results in intractable vocabulary sizes. For example, 252 languages $\times$ 32K tokens would result in a vocabulary size of 8.1M tokens, requiring 1.0B embedding parameters even with our smallest embedding size of 128. To avoid intractable parameter counts, we pre-train multilingual language models that each keep tokenization fixed for only one language, which we call the *target language* for that model. In each multilingual model, the non-target languages share a multilingual tokenizer with vocabulary size 32K, trained on 10K randomly-sampled lines from each added language. The target language and added multilingual datasets are tokenized separately, and the token IDs are merged for the shared vocabulary items. This merged tokenization process ensures that the target language tokenization remains unchanged across models.

Linguistic similarity. Motivated by work demonstrating the importance of linguistic similarity for crosslingual transfer performance (Pires et al., 2019; Conneau et al., 2020b; Gerz et al., 2018; Winata et al., 2022; Fujinuma et al., 2022; Ahuja et al., 2022; Imani et al., 2023; Oladipo et al., 2022; Eronen et al., 2023), we select added languages for multilingual data based on their similarity to each target language. Due to limits on computational resources, we only consider two linguistic similarity levels: similar and dissimilar languages.

Our linguistic similarity metric is based on three features: syntactic similarity, geographic proximity, and lexical similarity (i.e. tokenizer vocabulary overlap). Syntactic and geographic metrics are computed as cosine similarities between languages' syntactic and geographic vector representations from `lang2vec` (Littell et al., 2017), which pulls from the World Atlas of Language Structures (Dryer and Haspelmath, 2013). Lexical similarity is computed as the log number of shared tokens in the monolingual tokenizers for two languages (§4.1). We Z-score normalize the syntactic, geographic, and lexical similarity metrics over all language pairs, and we define the linguistic similarity between any two languages as the mean of the three similarity scores. For example, the four most similar languages to English are Dutch, Swedish, Norwegian, and German. The four most dissimilar languages to English are Nepali, Japanese, Tamil, and Korean. To allow us to vary the multilingual data quantity without changing the added languages, we restrict our added languages to those with at least 100M tokens in our dataset (i.e. our 48 med-high resource languages).

Manipulated variables. We manipulate four variables in our multilingual language modeling experiments:

- **Monolingual data quantity.** As in §4, we consider four monolingual dataset sizes when available in the target language: 1M, 10M, 100M, and 1B tokens.
- **Multilingual data quantity.** We always add multilingual data from 10 languages, selected according to linguistic similarity as described above. We add an equal number of tokens from each language, totaling either 10M, 100M, or 1B tokens. To save computation resources, we omit the 10M added tokens scenario when the monolingual data is 100M or 1B tokens.
- **Linguistic similarity.** When adding multilingual data for each target language, we select either the 10 most or 10 least similar languages

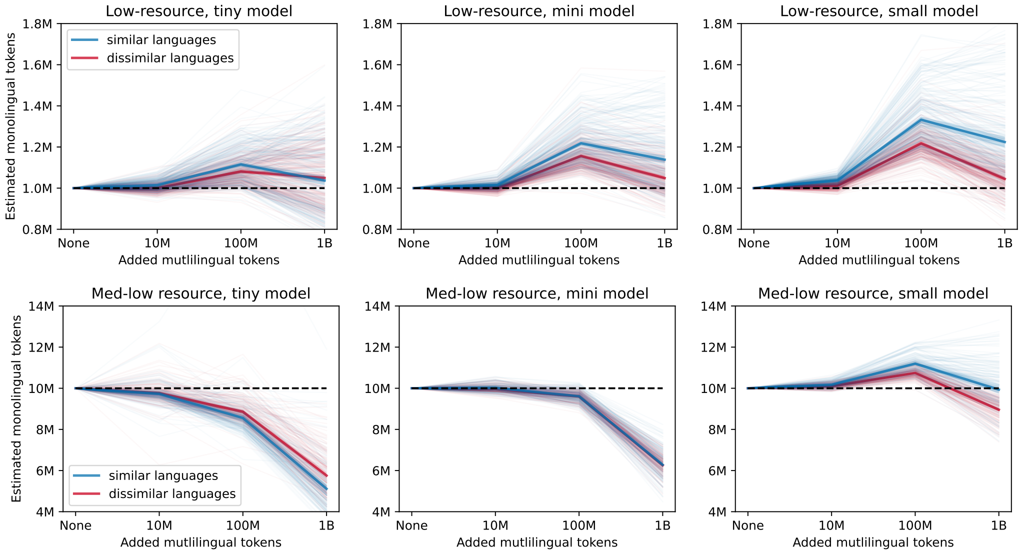

Figure 3: Results for low and med-low resource scenarios. Higher y-axis values indicate better performance. For example, a small model with 1M monolingual tokens (top right) and 1B added tokens of multilingual data in similar languages has similar performance to 1.2M monolingual tokens alone. Light-colored lines indicate results for individual languages, and bold lines indicate the mean across languages. Shaded regions are 95% confidence intervals for the mean.

using the similarity metric described above.

- **Model size.** We use the same model sizes as §4. With the added multilingual vocabulary embeddings, the models have roughly 8.7M (tiny), 19.8M (mini), and 45.8M (small) parameters.

6 Multilingual Model Results

To reflect low-resource to high-resource language scenarios, we primarily separate results based on monolingual data quantity (low and med-low resource in §6.1, high and med-high resource in §6.2). In each scenario, we consider the effects of adding different multilingual data quantities with different levels of linguistic similarity, across all three model sizes. Overall, we find that performance in low-resource languages improves when we add moderate amounts of multilingual data (§6.1). The amount of improvement depends on the syntactic similarity of the added languages, with small additional effects of lexical (vocabulary) similarity. High-resource language performance consistently degrades when we add multilingual data (§6.2). Larger models have smaller degradations for high-resource languages and larger improvements for low-resource languages in multilingual scenarios, suggesting that many drawbacks of multilinguality are due to limited model capacity.

6.1 Low-Resource Language Results

In moderation, multilinguality improves low-resource performance. As shown in Figure 3 (top), low-resource languages exhibit performance improvements when adding 100M or 1B tokens of multilingual data ($p < 0.001$ for 11 out of 12 comparisons, using paired sample t-tests; §E). Performance improvements are significantly larger when the added languages are similar vs. dissimilar to the target language (analogous to an average 33% vs. 22% increase in target language dataset size for small models in the optimal scenario; $p < 0.001$). Performance improvements are also larger for larger model sizes (33% vs. 12% equivalent dataset increases for small vs. tiny models; $p < 0.001$). Regardless of model size, performance is essentially unaffected when adding only 10M multilingual tokens (1M tokens in each added language); this result also holds for med-low resource scenarios (Figure 3, bottom). This suggests that a nontrivial amount of multilingual data is required for language models to leverage shared characteristics across languages.

However, the benefits of adding more multilingual data quickly plateau in low-resource scenarios (e.g. adding 100M vs. 1B multilingual tokens). In med-low resource scenarios (Figure 3, bottom), adding multilingual data hurts performance ($p < 0.001$ adding 1B multilingual tokens)

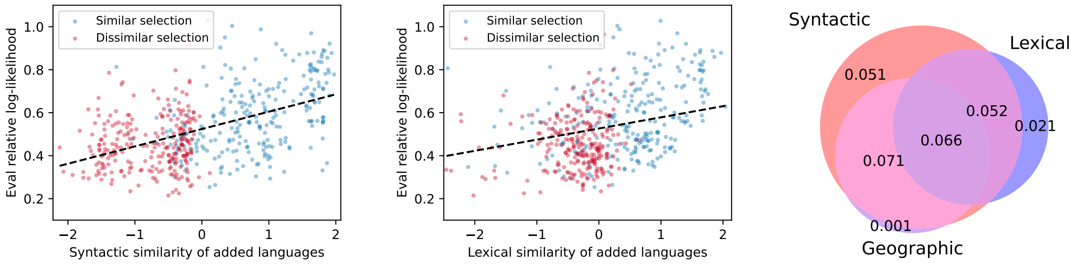

Figure 4: Left: Correlation between the mean syntactic similarity of the added languages and a model's relative log-likelihood score for the target language (Pearson's $r = 0.494$). Added languages are selected to be either similar or dissimilar (§5). A relative log-likelihood of 1.0 indicates that the model assigns the eval dataset $2^{1.0}$ times the likelihood assigned by the baseline model for that language. Center: Correlation ($r = 0.346$) between the mean lexical (vocabulary) similarity of the added languages and a model's relative log-likelihood score. Right: Variance partitioning into syntactic, geographic, and lexical similarity of the added languages when predicting a model's relative log-likelihood score. Additional results in §F.

except in our largest models. Even in the larger models, the benefits of multilinguality decrease when too much multilingual data is added (Figure 3, right). This suggests that adding multilingual data is beneficial only in moderation, before models have reached their capacity limits.

Syntactic similarity of added languages drives results. We then investigate whether syntactic, geographic, or lexical (vocabulary) similarity of the added languages appears to drive multilingual model improvement. We focus on the low-resource small model scenario (Figure 3, top right) with 100M tokens of added multilingual data. This setup leads to our largest performance improvement on average for low-resource languages; other scenarios are considered in §F. We compute the mean syntactic, geographic, and lexical similarity of the added languages for each target language, both when selecting languages based on similarity and dissimilarity. All three similarity metrics correlate with model performance (relative log-likelihood scores), with Pearson's $r = 0.494$, $r = 0.341$, and $r = 0.346$ respectively (Figure 4). For each type of similarity, more similar added languages correlate with better performance in the target language.

However, syntactic, geographic, and lexical similarity are also significantly correlated with one another ($r = 0.242$ to 0.602). We use variance partitioning to determine the amount of variance in model performance accounted for by each feature, along with the variance accounted for by each feature after regressing out the effects of other features (Borcard et al., 1992; QCBS, 2023). We find that syntactic similarity of the added languages accounts for 24.2% of variance in multilingual model performance; adding geographic and lexical similarity increases this to only 26.4% (Figure 4, right). We note that syntactic similarity might reflect other typological features of languages or be serving as a proxy for taxonomic relatedness (Rama and Kolachina, 2012). Still, these results suggest that abstract linguistic similarity drives the benefits of multilinguality more so than surface level features such as vocabulary overlap. This aligns with results for cross-lingual transfer during fine-tuning (Karthikeyan et al., 2020).

6.2 High-Resource Language Results

Multilinguality hurts high-resource performance. For all model sizes, multilinguality hurts language model performance in med-high and high resource languages (Figure 5; $p < 0.001$ in all scenarios adding 1B tokens, using paired sample t-tests; §E). For high-resource languages in our largest model size, adding 1B multilingual tokens is similar to removing 63% of the target language dataset. Degradations are larger when more multilingual tokens are added. Degradations are also larger for smaller models (88% vs. 63% equivalent dataset decrease in the target language for tiny vs. small models; $p < 0.001$). This suggests that degradations are likely driven by language models reaching capacity limits. Interestingly, degradations are slightly larger given more similar added languages to the target language (all scenarios in Figure 5; $p < 0.05$ in 7 out of 12 scenarios). This indicates that although more similar languages tend to improve low-resource language performance (§6.1), they surprisingly tend to hurt high-resource language performance more. One possible explanation is that more similar languages simply have larger effects on target language predictions. In

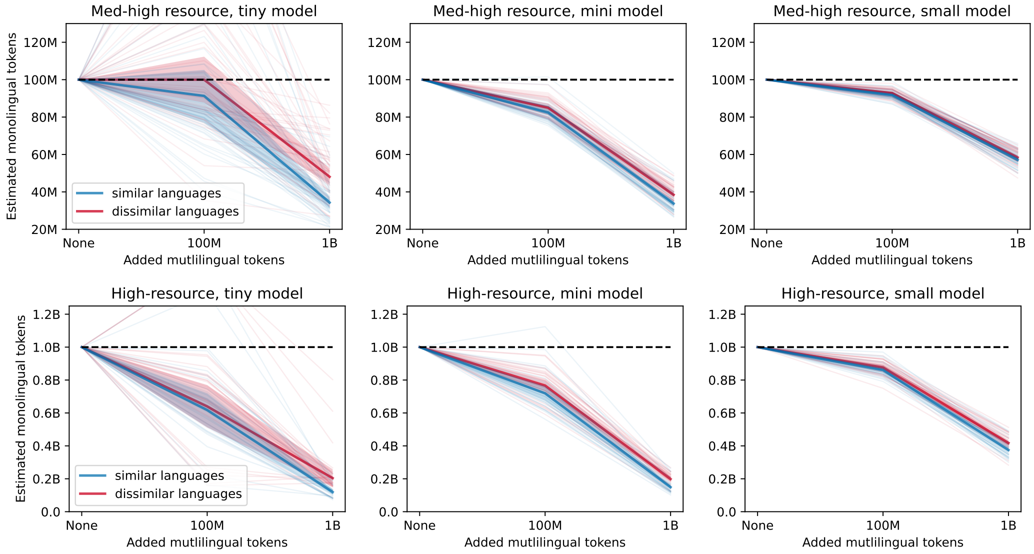

Figure 5: Results for med-high and high resource scenarios, using the same format as the low-resource scenarios in Figure 3. For example, adding 1B tokens of multilingual data to a small model with 1B monolingual tokens (high-resource; bottom right) is similar to removing over 600M tokens of the monolingual dataset.

low-resource scenarios, these influences from other languages improve predictions; however, in high-resource scenarios, when models are learning more fine-grained language-specific nuances, these influences might hurt performance, making similar languages hurt performance more.

7 Discussion

Our results demonstrate that for low-resource languages, multilingual language models yield some benefits. In the optimal case from our study, the benefits are similar to increasing the low-resource dataset size by about 33% (§6.1). Hence, in scenarios where collecting additional data is difficult (e.g. languages spoken in remote geographic locations or with few speakers), pre-training multilingual models may be a worthwhile endeavor. In these cases, the models should be pre-trained with multilingual data from maximally similar languages, and it should be ensured that the models have capacity for the added multilingual data along with the target language data. However, in other cases, it may be more practical to find or collect more data in the target language itself (e.g. if collecting 50% more target language data is feasible).

For high-resource languages, multilingual language models yield worse performance than the comparable monolingual model in essentially all cases. Degradations can be similar to reducing high-resource dataset sizes by over 85% (§6.2). These degradations can be mitigated by pre-training larger models, which also appear to maximize benefits for low-resource languages. However, when pre-training language models even on the order of tens of high-resource languages (Conneau et al., 2020a; Scao et al., 2022; Lin et al., 2022), a model sufficiently large to accommodate all of the languages' data without hitting capacity limitations would likely be impractically large. Even if existing language models are severely over-parameterized, there is evidence that 70B-parameter models are required just for English (Hoffmann et al., 2022). If only considering individual language performance, pre-training targeted language-specific models is likely to be far more efficient than a single massively multilingual model.

8 Conclusion

Our work systematically evaluates the effects of multilingual pre-training on language modeling performance in 252 languages. We pre-train over 10,000 monolingual and multilingual language models, varying monolingual dataset sizes, multilingual dataset sizes, linguistic similarity of the multilingual data, and model sizes. We find that adding multilingual data in similar languages improves performance for low-resource languages, but improvements decrease as models reach capacity limitations. Multilingual data consistently hurts high-resource language performance. We quantify both of these effects in terms of comparable monolingual dataset sizes. Our results suggest that while multilingual language models may be beneficial for low-resource scenarios, massively multilingual models may be far less practical than previously assumed for raw language modeling.

Limitations

Model size. We only pre-train language models up to 45M parameters. Larger models are less likely to hit the capacity limitations that appear to drive the "curse of multilinguality", but we select our model sizes as a compromise between informativity of results and computational cost. When pre-training thousands of models for controlled experiments, larger models may not be worth additional computational and environmental costs if results can reasonably be extrapolated to larger models (Strubell et al., 2019). In our experiments, directions of effect are consistent across all three model sizes we evaluate.

In fact, for low-resource scenarios, smaller models can achieve similar performance to larger models (Figure 2) while remaining accessible to communities with fewer computational resources. This makes small models useful for efficient low-resource language technologies and low-compute settings such as laptops and mobile phones. Pre-training smaller models in our experiments also allows us to include a much larger and more typologically diverse set of languages in our study, making our results more representative of human languages overall and more likely to generalize to languages not included in our study. Our results are much less likely to be skewed by over-representation of the small number of languages that dominate the field of NLP (Joshi et al., 2020; Blasi et al., 2022).

Language coverage. While we have included far more low-resource languages than the vast majority of recent studies in NLP, we do not have coverage of some regions and language families. For example, our study does not include any languages indigenous to modern-day Australia or many from the Americas. This imperfect coverage may skew our results towards languages that have larger text corpora available on the Internet. Specifically, as discussed in §5, because we restrict added multilingual data to our 48 med-high resource languages (to allow us to vary multilingual dataset sizes), our low-resource target languages are less likely to have highly similar languages in the multilingual pre-training scenarios. Allowing added multilingual data from our low and med-low resource languages would increase the mean similarity of added similar languages in §6.1, so we would expect to see larger performance improvements for low-resource languages in these cases (i.e. the observed equivalent 33% dataset increase for low-resource languages would likely be greater); this can be tested empirically in future work. Our work demonstrates that low-resource performance improvements can be predicted by the syntactic similarity of added languages (moreso than lexical overlap; §6.1), but future research might investigate more specific syntactic and semantic features that result in high crosslingual transfer.

Of course, as with all multilingual datasets, it is likely that there are still some language labeling mismatches and contaminated examples in our dataset. We also note that the delineations defining languages versus different dialects of the same language are inherently fuzzy. For example, Northern Frisian (frr), Eastern Frisian (frs), and Western Frisian (fry) are considered individual languages with separate codes; conversely, Tamil (tam) is considered an individual language (one language code) with at least 18 dialects (Hammarström et al., 2023). We defer to the ISO 639-3 language code system, as it is the most widely used system of its type.

Measuring performance. Finally, our results apply primarily to language modeling performance. Effects of multilingual pre-training may be different for specific downstream tasks (e.g. reasoning tasks or machine translation; Bandarkar et al., 2023; Costa-jussà et al., 2022) or for cross-lingual transfer learning using fine-tuning (Fujinuma et al., 2022). Unfortunately, few existing multilingual benchmarks cover the wide variety of languages used in our study. There are several evaluations used in Imani et al. (2023); however, with the exception of perplexity, all of the Glot500 evals are designed primarily for bidirectional models, or they evaluate crosslingual performance rather than a single target language: sentence retrieval, Bible text classification, NER, POS tagging, and roundtrip alignment. Bidirectional (encoder) models remain quite useful for representation learning (e.g. sentence representation and classification tasks; Bandarkar et al., 2023; Imani et al., 2023; Conneau et al., 2020a), but the majority of recent language model training efforts have focused on autoregressive models (e.g. XGLM and BLOOM, along with multilingual capabilities of GPT-4, Claude, Gemini, etc). To best align our work with current pre-training efforts, we focus on autoregressive models.

Of the datasets that do exist for low-resource language evaluation, Belebele is a massively multilingual reading comprehension dataset, which covers only 122 language variants (Bandarkar et al., 2023), and the XTREME benchmark covers only 40 languages (Hu et al., 2020), all of which are at least medium-low resource (i.e. not low-resource) in our study. We use evaluation log-likelihoods (negative log-perplexities) to measure language modeling performance in our experiments in order to evaluate all the languages in our sample with the same metric. Evaluation log-likelihoods require no annotated data in the target language, they are predictive of language model behavior on a variety of tasks (Xia et al., 2023), and they have been used to quantify language model quality in previous work (Kaplan et al., 2020; Hoffmann et al., 2022). As multilingual language models are increasingly used without fine-tuning for raw text generation (e.g. Scao et al., 2022; Lin et al., 2022; OpenAI, 2023; Google DeepMind, 2023), raw language modeling performance across languages is increasingly important to evaluate.

Acknowledgments

We would like to thank the UCSD Language and Cognition Lab for valuable discussion. Some models were trained on hardware provided by the NVIDIA Corporation as part of an NVIDIA Academic Hardware Grant. Some models were also trained on the UCSD Social Sciences Research and Development Environment (SSRDE). Zhuowen Tu is supported by NSF IIS-2127544. Tyler Chang is partially supported by the UCSD HDSI graduate fellowship.

References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event)*, pages 1 – 9.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Kabir Ahuja, Sunayana Sitaram, Sandipan Dandapat, and Monojit Choudhury. 2022. On the calibration of massively multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4310–4323. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349. International Committee on Computational Linguistics.

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

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

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

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505. Association for Computational Linguistics.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3575–3590. Association for Computational Linguistics.

C.E. Bonferroni. 1936. Teoria statistica delle classi ecalcolo delle probabilità. *Pubblicazioni del R Instituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62.

Daniel Borcard, Pierre Legendre, and Pierre Drapeau. 1992. Partialling out the spatial component of ecological variation. *Ecology*, 73(3):1045–1055.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Cawoylel. 2023. Fula speech corpus.

Yuan Chai, Yaobo Liang, and Nan Duan. 2022. Cross-lingual ability of multilingual masked language models: A study of language structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4702–4712. Association for Computational Linguistics.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Cherokee Corpus. 2023. Cherokee corpus and Cherokee-English Dictionary.

CMU. 2010. Haitian Creole language data. `http://www.speech.cs.cmu.edu/haitian/`.

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

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.3)*. Zenodo.

eBible. 2023. eBible.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299. Association for Computational Linguistics.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250.

Yoshinari Fujinuma, Jordan Boyd-Graber, and Katharina Kann. 2022. Match the script, adapt if multilingual: Analyzing the effect of multilingual pretraining on cross-lingual transferability. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1500–1512. Association for Computational Linguistics.

Fitsum Gaim, Wonsuk Yang, and Jong Park. 2021. Monolingual pre-trained language models for Tigrinya. *Widening NLP Workshop (WiNLP)*.

Yvette Gbedevi Akouyo, Kevin Zhang, and Tchaye-Kondi Jude. 2021. GELR: A bilingual Ewe-English corpus building and evaluation. *International Journal of Engineering Research and Technology (IJERT)*, 10.

Daniela Gerz, Ivan Vulić, Edoardo Maria Ponti, Roi Reichart, and Anna Korhonen. 2018. On the relation between linguistic typology and (limitations of) multilingual language modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 316–327. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765. European Language Resources Association (ELRA).

Google DeepMind. 2023. Gemini: A family of highly capable multimodal models. *Google*.

Raffaele Guarasci, Stefano Silvestri, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2022. BERT syntactic transfer: A computational experiment on Italian, French and English languages. *Computer Speech & Language*, 71:101261.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. *Glottolog 4.8*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python. SpaCy.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117. Association for Computational Linguistics.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. 2020. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961. Association for Computational Linguistics.

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

Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: An empirical study. In *International Conference on Learning Representations*.

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

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8424–8445. Association for Computational Linguistics.

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. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217. Association for Computational Linguistics.

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

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126. Association for Computational Linguistics.

Tolulope Ogunremi, Dan Jurafsky, and Christopher Manning. 2023. Mini but mighty: Efficient multilingual pretraining with linguistically-informed data selection. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1251–1266.

Akintunde Oladipo, Odunayo Ogundepo, Kelechi Ogueji, and Jimmy Lin. 2022. An exploration of vocabulary size and transfer effects in multilingual language models for African languages. In *3rd Workshop on African Natural Language Processing*.

OpenAI. 2023. GPT-4 technical report. *OpenAI*.

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

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001. Association for Computational Linguistics.

Kholisa Podile and Roald Eiselen. 2016. NCHLT isiXhosa Named Entity Annotated Corpus.

Sampo Pyysalo, Jenna Kanerva, Antti Virtanen, and Filip Ginter. 2021. WikiBERT models: Deep transfer learning for many languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 1–10, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

QCBS. 2023. Advanced Multivariate Analyses in R: Variation Partitioning. In *QCBS R Workshop Series*. Québec Centre for Biodiversity Science.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*.

Taraka Rama and Prasanth Kolachina. 2012. How good are typological distances for determining genealogical relationships among languages? In *Proceedings of COLING 2012*, pages 975–984. The COLING 2012 Organizing Committee.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Elizabeth-Jane Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Francois Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Rose Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650. Association for Computational Linguistics.

Daniela Teodorescu, Josie Matalski, Delaney Lothian, Denilson Barbosa, and Carrie Demmans Epp. 2022. Cree corpus: A collection of nêhiyawêwin resources.

In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6354–6364. Association for Computational Linguistics.

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

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv*.

Ulukau. 2023. Ulukau: The Hawaiian Electronic Library. `https://ulukau.org/index.php?l=en`.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450. Association for Computational Linguistics.

Wikimedia. 2023. Wikimedia dumps.

Wikipedia. 2023. Wikipedia.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 777–791. Association for Computational Linguistics.

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

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130. Association for Computational Linguistics.

Zhengxuan Wu, Isabel Papadimitriou, and Alex Tamkin. 2022. Oolong: Investigating what makes crosslingual transfer hard with controlled studies. *arXiv*.

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Veselin Stoyanov. 2023. Training trajectories of language models across scales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13711–13738. Association for Computational Linguistics.

Lyudmila Zaydelman, Irina Krylova, and Boris Orekhov. 2016. The technology of web-texts collection of Russian minor languages. In *Proceedings of the International Scientific Conference CPT2015*, pages 179–181.

Shiyue Zhang, Benjamin Frey, and Mohit Bansal. 2020. ChrEn: Cherokee-English machine translation for endangered language revitalization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 577–595. Association for Computational Linguistics.

Anna Zueva, Anastasia Kuznetsova, and Francis Tyers. 2020. A finite-state morphological analyser for Evenki. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2581–2589. European Language Resources Association.

A Dataset Details

We first download the first 32M lines for each language in the deduplicated September 2021 release of OSCAR (Ortiz Suárez et al., 2019; Abadji et al., 2021). We collect additional corpora for languages with less than 1M lines in OSCAR (approximately 50M tokens based on OSCAR line lengths) and for languages that do not appear in OSCAR. Additional corpora consist of: Wikipedia (Wikipedia, 2023), NLLB (Costa-jussà et al., 2022), the Leipzig Corpora Collection (Goldhahn et al., 2012), eBible translations (eBible, 2023), FLORES-200 (Costa-jussà et al., 2022), Tatoeba (Tiedemann, 2012, 2020), AfriBERTa (Ogueji et al., 2021), NusaX (Winata et al., 2023), AmericasNLP (Mager et al., 2021), AmericasNLI (Ebrahimi et al., 2022), the Nunavut Hansard Inuktitut–English Parallel Corpus (Joanis et al., 2020), the Cherokee-English ChrEn dataset (Zhang et al., 2020), the Cherokee Corpus (Cherokee Corpus, 2023), the Cree Corpus (Teodorescu et al., 2022), Languages of Russia (Zaydelman et al., 2016), the Evenki Life newspaper (Zueva et al., 2020), the transcribed Fula Speech Corpora (Cawoylel, 2023), IsiXhosa (Podile and Eiselen, 2016), the Ewe Language Corpus (Gbedevi Akouyo et al., 2021), the Makerere Luganda Corpora (Mukiibi et al., 2022), the CMU Haitian Creole dataset (CMU, 2010), the Tigrinya Language Modeling Dataset (Gaim et al., 2021), and Ulukau (Ulukau, 2023). Our Wikipedia corpora use the Wikimedia dump from August 20, 2023 (Wikimedia, 2023). All other corpora use their publicly available versions as of August 2023. Links to individual corpora are included at `https://github.com/tylerachang/curse-of-multilinguality`. While we are unable to redistribute our compiled dataset due to redistribution licenses and out of respect for the original data collectors, all of our sources are publicly available. As a caveat, we note that many low-resource language datasets prohibit commercial use, and thus industry labs may be precluded from using such datasets without explicit permission from the owners.

We clean each corpus by removing lines consisting of only repetitive characters, exact duplicate lines, and lines identified as English by the `spaCy` language detection tool with confidence above 0.95 (except for the English dataset; Honnibal et al., 2020). We find that English filtering is particularly important for Wikipedia, from which we also remove redundant lists of links and headers. We manually inspect all files for egregious unclean text lines, and we remove any patterns found. All corpora outside of OSCAR are truncated to 2M cleaned lines per language, which encompasses the entire corpus for most datasets; for example, only 4 out of 239 downloaded Wikipedias are truncated (recall that we only download additional corpora for languages with less than 1M lines in OSCAR). Corpora are unshuffled unless their public release is already shuffled. This allows tokenized sequences to span multiple consecutive lines; the tokenized sequences are shuffled prior to language model pre-training. Final token counts per language are listed in §G.

B Tokenizer Details

To control for tokenization quality across languages, all of our monolingual tokenizers are SentencePiece tokenizers trained on 10K lines of text with maximum vocabulary size 32K (§4.1; Kudo and Richardson, 2018). We have at least 10K lines of text in each of our 252 languages. All evaluations (including for multilingual models, which fix the target language monolingual tokenizer) are conducted using these tokenizers. The multilingual tokenizers in §5 are used only for added data during multilingual pre-training; they are not used for evaluation. To ensure that our monolingual tokenizers have reasonable quality, we compare their vocabularies with tokenizers trained on more lines of text. Specifically, for each of our 28 high-resource languages, we train tokenizers on 10K, 100K, 1M, and 10M lines of text. For each training dataset size, we compute the vocabulary overlap with the 4K and 8K most frequent tokens in the 10M-line tokenizer (the "reference vocabulary"). Figure 6 shows the reference vocabulary overlap for the different training dataset sizes. At 10K lines, the tokenizer vocabularies on average cover 93.7% of the 4K-token reference vocabulary and 87.8% of the 8K-token reference vocabulary, indicating reasonable tokenization quality.

C Language Model Pre-Training Details

Language models are pre-trained using the Hugging Face Transformers library (Wolf et al., 2020) and code from Chang and Bergen (2022). Hyperparameters are reported in Table 1 (left). All of our models use the GPT-2 architecture (Radford et al., 2019), changing only the number of layers, atten-

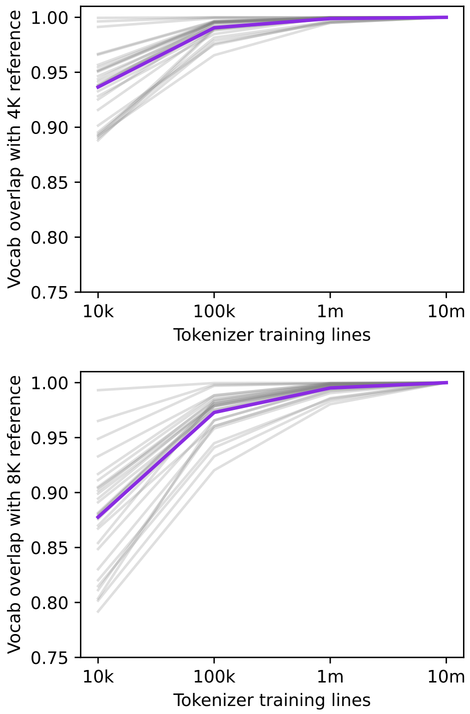

Figure 6: Vocabulary overlap with the reference vocabulary for tokenizers trained on different numbers of lines. The reference vocabulary consists of the 4K (left) or 8K (right) most frequent tokens in a 10M-line tokenizer for that language. We report the proportion of the reference vocabulary that is covered by 32K-vocabulary tokenizers with different training dataset sizes. Gray lines indicate individual languages, and the purple line indicates the mean across languages.

tion heads, and embedding sizes as in Turc et al. (2019). Models are pre-trained for 20 epochs of the target language monolingual data in the low and med-low resource scenarios, 10 epochs in the med-high resource scenario, and 2 epochs in the high-resource scenario. Based on initial results using randomly-sampled languages, pre-training on more than 20 epochs often leads to overfitting (increases in eval loss) in low-resource scenarios. Multilingual models include one epoch of the multilingual data (§5) randomly interspersed with the target language data. The numbers of pre-training steps for different dataset configurations are reported in Table 1 (right). Average evaluation loss curves during pre-training are shown in Figure 7. For each target language, the same 500K evaluation tokens are held out in all cases. In the monolingual low-resource scenario for each language (i.e. 1M pre-training tokens), we pre-train three tiny models (instead of one) and compute their average evaluation log-likelihood, because these models are used as the baseline models for relative log-likelihoods (§4.2).

All language model pre-training runs together take a total of 1.87×10^{20} FLOPs. This is less than $1/1500\times$ the computation used to train the original 175B-parameter GPT-3 model (Brown et al., 2020; 3.14×10^{23} FLOPs). Models are each trained on one NVIDIA GeForce GTX TITAN X, GeForce RTX 2080 Ti, TITAN Xp, Quadro P6000, RTX A4500, RTX A5000, or RTX A6000 GPU. Our pre-training experiments take approximately 17700 A6000 GPU hours. Dataset cleaning, tokenization, and merging takes approximately 5880 CPU core hours, largely due to dataset tokenization with each multilingual tokenizer.

D Monolingual Token Estimation Details

We overview our monolingual token estimation process in §4.3, and we provide details here. As motivation, we note that relative log-likelihood scores are not comparable across model sizes. For example, suppose that adding a multilingual dataset D improves a model's eval log-likelihood score by 1.0 in both small and large models. In this case, it would be unclear whether the effect of D is intuitively "equal" in the two model sizes; doubling the likelihood of the eval dataset is likely more difficult in the larger model, so we might interpret D as having a larger effect on the larger model despite the same change in log-likelihood. To avoid this ambiguity, we measure model performance using the estimated number of monolingual tokens in the target language that would achieve similar performance. In the case above, adding the multilingual dataset D might be similar to adding n_1 monolingual tokens to the smaller model, but similar to adding $n_2 > n_1$ monolingual tokens to the larger model.

To estimate this, we first fit a power law $-ax^{-b}+c$ for each of our 252 languages, predicting a model's relative log-likelihood score (§4.2) based on its pre-training dataset size in log10 tokens. Each language has up to four ground truth values, corresponding to our monolingual models pre-trained on 1M, 10M, 100M, and 1B tokens. When all four points are available (i.e. our 28 high-resource languages), we are able to fit a power law from scratch. From these languages, we estimate the medians and standard deviations of a, b, and c. For languages with fewer than four data points, we constrain a, b, and c to be within 2.5 standard deviations from the median parameter value. If this

Hyperparameter	Tiny	Mini	Small
Layers	2	4	4
Embedding size	128	256	512
Hidden size	128	256	512
Intermediate hidden size	512	1024	2048
Attention heads	2	4	8
Attention head size	64	64	64
Learning rate	1e-3	7e-4	5e-4
Activation function	GELU		
Max sequence length	128		
Position embedding	Absolute		
Batch size	128		
Learning rate decay	Linear		
Warmup steps	10% of pre-training		
Adam ϵ	1e-6		
Adam β_1	0.9		
Adam β_2	0.999		
Dropout	0.1		
Attention dropout	0.1		

Mono. tokens	Mono. epochs	Multi. tokens	Pre-training steps
1M	20	0	1250
1M	20	10M	1875
1M	20	100M	7500
1M	20	1B	63750
10M	20	0	12500
10M	20	10M	13125
10M	20	100M	18750
10M	20	1B	75000
100M	10	0	62500
100M	10	100M	68750
100M	10	1B	125000
1B	2	0	125000
1B	2	100M	131250
1B	2	1B	187500

Table 1: Left: Language model pre-training hyperparameters (Devlin et al., 2019; Turc et al., 2019; Radford et al., 2018). To prevent overfitting (increasing loss on the eval dataset), learning rates are halved for mini and small models in the low-resource scenario, to 4e-4 and 2e-4 respectively (§4.1). Right: Pre-training steps for different monolingual and multilingual dataset sizes. There is always one epoch of the multilingual dataset (§5).

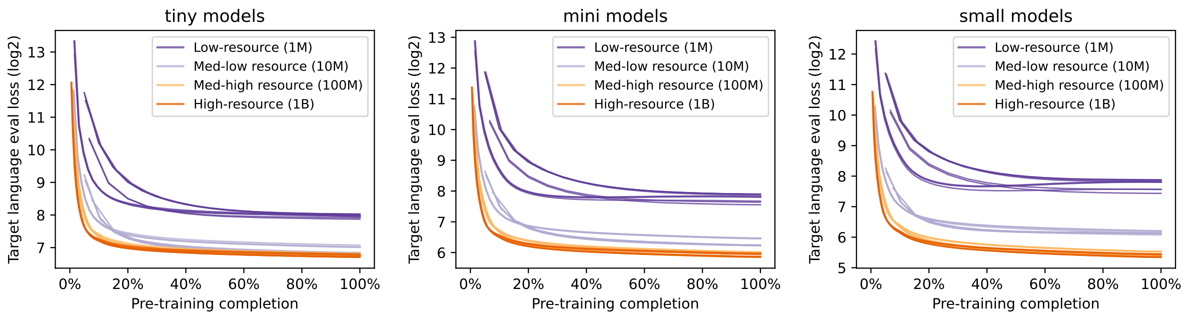

Figure 7: Target language evaluation loss curves during pre-training, for different model sizes and language resource scenarios. Each individual curve corresponds to a dataset configuration in Table 1 (right), averaging the loss curve over languages.

leads the curve fitting to diverge, we loosen this constraint to 5.0, 7.5, then 10.0 standard deviations from the median.

For languages where the curve fitting still does not converge or languages with too few data points (e.g. med-low resource languages with data points only for 1M and 10M tokens), we fix a as the median parameter value from the high-resource languages. We fit only b and c, which we constrain using standard deviations in the same way as described above. If the curve fitting still does not converge when fixing a (e.g. low-resource languages with a data point only for 1M tokens), we fix both a and b as their median values. In that case, we only fit c, which is equivalent to simply shifting the median curve up or down by a constant. All curve fitting is implemented using `scipy` (Virtanen et al., 2020).

Finally, in many cases, we compare multilingual models to monolingual models with a specific known dataset size. The multilingual models in §6 are all compared to corresponding monolingual models without any added multilingual data. For example, a multilingual model with 10M monolingual tokens and 100M added multilingual tokens (relative log-likelihood score y_1) would be compared to a monolingual model with 10M monolingual tokens alone (relative log-likelihood score y_0). In these cases, we constrain our curve-fitting to pass through the point corresponding to the reference monolingual model (e.g. in the example described, the curve would be required to pass through the ground truth point $(7.0, y_0)$ for $10^{7.0}$ monolingual tokens alone). This only slightly alters the curve predicting relative log-likelihood score from log10 tokens, but it ensures that our baseline monolingual models in §6 lie exactly at 1M, 10M, 100M, and 1B tokens (Figure 3 and Figure 5).

Once we have fitted a curve predicting a model's relative log-likelihood score from log10 pre-training tokens in a language L, we use this curve to estimate the number of tokens required to achieve any relative log-likelihood score. Then, we have two metrics for a multilingual model's performance on target language L: (1) the model's relative log-likelihood score itself and (2) the estimated number of monolingual tokens in L that would achieve that relative log-likelihood. The latter metric is easily interpretable, and it facilitates comparisons across languages and model sizes. We note that the estimated token count is a monotonic increasing function of relative log-likelihood score in all cases. Thus, even if the estimated token counts are not perfectly accurate, they preserve performance rankings between models (e.g. between our multilingual models and the monolingual baselines). A language model with target language L will have a higher estimated token count if and only if it assigns a higher log-likelihood score to the evaluation dataset for L.

Still, we evaluate the quality of our monolingual token count estimation process. For each language L, we have up to four monolingual models (1M, 10M, 100M, and 1B pre-training tokens). We hold out one (or multiple) of the models, and we estimate its monolingual token count based on a curve fitted to the other monolingual models for L. We note that these estimations are extrapolating at minimum one order of magnitude away from the models used to fit the curve, because the models are exactly one order of magnitude apart in terms of pre-training tokens. The results in §6 do not need to extrapolate this far. Still, even with this larger extrapolation, we obtain reasonable estimates of monolingual token counts in the held-out scenarios (Figure 8). The root-mean-square errors are 0.340, 0.317, and 0.335 log10 tokens for tiny, mini, and small models respectively. Again, regardless of estimation quality, the estimated token counts are simply a monotonic increasing function of relative log-likelihood score.

E Statistical Tests

We run paired sample t-tests to assess the statistical significance of our results from §6. For each reported p-value, we compare models that differ by exactly one of: monolingual dataset size, multilingual dataset size, linguistic similarity of the added languages, or model size. We pair models by language, so each pair differs by only the manipulated variable. To avoid potential artifacts from our token estimation process, we compare model relative log-likelihoods directly (§4.2) unless comparing across two model sizes (because relative log-likelihood improvements and degradations are difficult to compare across model sizes; §D). If comparing across model sizes, we compare the estimated monolingual token counts of the models. In both cases, we use a paired sample t-test. To decrease the chance of false positive results, we only run the statistical tests whose p-values are reported in the main text, and we account for multiple comparisons using Bonferroni correction (Bonferroni, 1936). For estimates of significance, the plots in §6 also include 95% confidence intervals for means.

F Additional Correlations

In §6.1, we find that the mean syntactic similarity of the added languages accounts for more variance in multilingual model performance (relative log-likelihood scores) than geographic and lexical (vocabulary) similarity. In that section, we consider the low-resource scenario with 100M added multilingual tokens in small models. Here, we report the same results for tiny, mini, and small models. Variance partitioning results are shown in Figure 9. In all cases, syntactic similarity accounts for more variance than geographic and lexical similarity. Correlations between different similarity measures and model performance for mini and tiny models with 100M added multilingual tokens are plotted in Figure 10.

G List of Languages

The 252 languages included in our language modeling study are listed in Table 2. These languages are those with at least 1.5M tokens in our dataset (§A). We restrict all languages to a maximum of 1B tokens. In lower resource scenarios, higher resource languages are subsampled to mimic the lower resource scenario. For example, we have 167 med-low resource languages when including the subsampled med-high and high resource languages. We

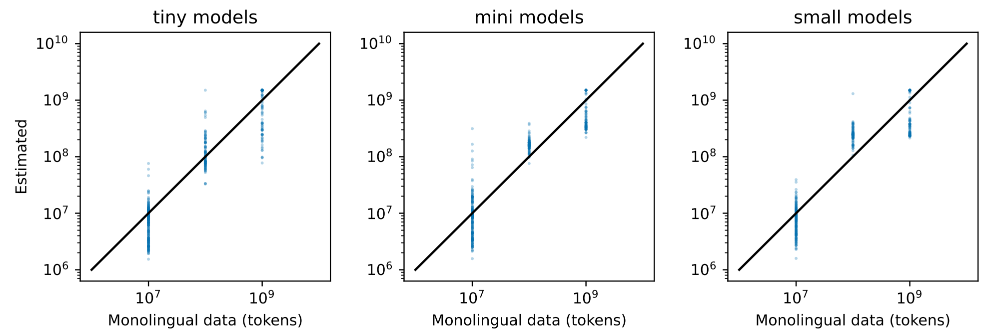

Figure 8: Estimated monolingual token counts for held-out monolingual models. Token counts are estimated from each model's relative log-likelihood score using a curve fitted to the specific language (§4.3). Estimations are extrapolating one order of magnitude out from the points used to fit the curve. In practice, we generally do not need to extrapolate this far for our results. The black line indicates perfect accuracy.

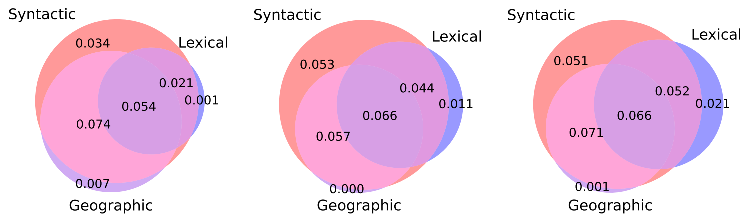

Figure 9: Variance partitioning into syntactic, geographic, and lexical similarity of the added languages when predicting a model's performance (relative log-likelihood score) for tiny (left), mini (center), and small (right) models with 100M tokens of added multilingual data.

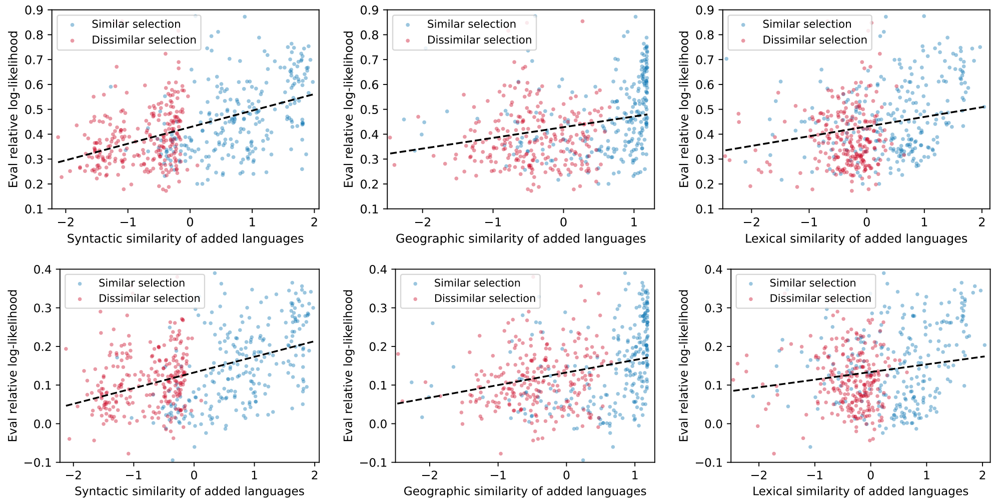

Figure 10: Top: Correlations between syntactic ($r = 0.471$), geographic ($r = 0.305$), and lexical ($r = 0.306$) similarity of added languages and target language performance for mini models, as described in §6.1. Bottom: Correlations between syntactic ($r = 0.430$), geographic ($r = 0.345$), and lexical ($r = 0.233$) similarity of added languages and target language performance for tiny models.

distinguish between the same language in multiple scripts (e.g. Serbian in Cyrillic vs. Latin script) and macrolanguages vs. their individual constituent languages (e.g. Quechua vs. Cusco Quechua and Ayacucho Quechua). The full list of 1572 languages in our dataset can be found at https://github.com/tylerachang/curse-of-multilinguality.

	Language	Language (ISO 639-3)	Script (ISO 15924)	Tokens	Resource Category	Language Family
1	Bulgarian	bul	cyrl	1024512000	high	Indo-European
2	Chinese	zho	hans	1024512000	high	Sino-Tibetan
3	Czech	ces	latn	1024512000	high	Indo-European
4	Danish	dan	latn	1024512000	high	Indo-European
5	Dutch	nld	latn	1024512000	high	Indo-European
6	English	eng	latn	1024512000	high	Indo-European
7	Finnish	fin	latn	1024512000	high	Uralic
8	French	fra	latn	1024512000	high	Indo-European
9	German	deu	latn	1024512000	high	Indo-European
10	Hebrew	heb	hebr	1024512000	high	Afro-Asiatic
11	Hungarian	hun	latn	1024512000	high	Uralic
12	Indonesian	ind	latn	1024512000	high	Austronesian
13	Iranian Persian	pes	arab	1024512000	high	Indo-European
14	Italian	ita	latn	1024512000	high	Indo-European
15	Japanese	jpn	jpan	1024512000	high	Japonic
16	Korean	kor	hang	1024512000	high	Koreanic
17	Modern Greek	ell	grek	1024512000	high	Indo-European
18	Polish	pol	latn	1024512000	high	Indo-European
19	Portuguese	por	latn	1024512000	high	Indo-European
20	Romanian	ron	latn	1024512000	high	Indo-European
21	Russian	rus	cyrl	1024512000	high	Indo-European
22	Spanish	spa	latn	1024512000	high	Indo-European
23	Standard Arabic	arb	arab	1024512000	high	Afro-Asiatic
24	Swedish	swe	latn	1024512000	high	Indo-European
25	Thai	tha	thai	1024512000	high	Tai-Kadai
26	Turkish	tur	latn	1024512000	high	Turkic
27	Ukrainian	ukr	cyrl	1024512000	high	Indo-European
28	Vietnamese	vie	latn	1024512000	high	Austro-Asiatic
29	Lithuanian	lit	latn	787855616	medhigh	Indo-European
30	Hindi	hin	deva	774095488	medhigh	Indo-European
31	Catalan	cat	latn	771223680	medhigh	Indo-European
32	Slovak	slk	latn	746472192	medhigh	Indo-European
33	Norwegian Bokmål	nob	latn	612469888	medhigh	Indo-European
34	Estonian	est	latn	500367232	medhigh	Uralic
35	Bengali	ben	beng	419860608	medhigh	Indo-European
36	Latvian	lav	latn	379466368	medhigh	Indo-European
37	Serbian	srp	cyrl	279173376	medhigh	Indo-European
38	Slovenian	slv	latn	270027392	medhigh	Indo-European
39	Tamil	tam	taml	257684608	medhigh	Dravidian
40	Albanian	sqi	latn	240805504	medhigh	Indo-European
41	Azerbaijani	aze	latn	178155008	medhigh	Turkic
42	Urdu	urd	arab	143181312	medhigh	Indo-European
43	Nepali	npi	deva	139989120	medhigh	Indo-European
46	Macedonian	mkd	cyrl	124803328	medhigh	Indo-European
47	Kazakh	kaz	cyrl	124020480	medhigh	Turkic
48	Georgian	kat	geor	122249472	medhigh	Kartvelian
49	Armenian	hye	armn	121111040	medhigh	Indo-European
50	Belarusian	bel	cyrl	108812544	medhigh	Indo-European
44	Esperanto	epo	latn	102911872	medlow	Constructed
45	Croatian	hrv	latn	102911872	medlow	Indo-European
51	Malayalam	mal	mlym	90062848	medlow	Dravidian
52	Icelandic	isl	latn	88493056	medlow	Indo-European
53	Welsh	cym	latn	86114176	medlow	Indo-European
54	Telugu	tel	telu	81769088	medlow	Dravidian
55	Galician	glg	latn	81455616	medlow	Indo-European
56	Hausa	hau	latn	81195520	medlow	Afro-Asiatic
57	Mongolian	mon	cyrl	79270528	medlow	Mongolic
58	Marathi	mar	deva	78900992	medlow	Indo-European
59	Asturian	ast	latn	76998272	medlow	Indo-European
60	Afrikaans	afr	latn	75925632	medlow	Indo-European
61	Basque	eus	latn	75490304	medlow	Basque
62	Burmese	mya	mymr	75295104	medlow	Sino-Tibetan
63	Bosnian	bos	latn	73321472	medlow	Indo-European
64	Central Kanuri	knc	arab	72147840	medlow	Nilo-Saharan

65	Somali	som	latn	71963648	medlow	Afro-Asiatic
66	Tatar	tat	cyrl	71448448	medlow	Turkic
67	Cebuano	ceb	latn	71133568	medlow	Austronesian
68	Kannada	kan	knda	69977600	medlow	Dravidian
69	Central Khmer	khm	khmr	67915392	medlow	Austro-Asiatic
70	Gujarati	guj	gujr	65388416	medlow	Indo-European
71	Panjabi	pan	guru	64354560	medlow	Indo-European
72	Bashkir	bak	cyrl	64024832	medlow	Turkic
73	Central Kurdish	ckb	arab	60765440	medlow	Indo-European
74	Maltese	mlt	latn	59164544	medlow	Afro-Asiatic
75	Serbo-Croatian	hbs	latn	58518784	medlow	Indo-European
76	Tajik	tgk	cyrl	57351424	medlow	Indo-European
77	Tagalog	tgl	latn	55507456	medlow	Austronesian
78	Kirghiz	kir	cyrl	55496576	medlow	Turkic
79	Tigrinya	tir	ethi	55378816	medlow	Afro-Asiatic
80	Malay	msa	latn	55249152	medlow	Austronesian
81	Igbo	ibo	latn	53409920	medlow	Niger-Congo
82	Sinhala	sin	sinh	53101952	medlow	Indo-European
83	Irish	gle	latn	51020544	medlow	Indo-European
84	Amharic	amh	ethi	49825536	medlow	Afro-Asiatic
85	Uzbek	uzb	latn	49750144	medlow	Turkic
86	Swahili	swa	latn	49580928	medlow	Atlantic-Congo
87	Luxembourgish	ltz	latn	46355968	medlow	Indo-European
88	Yoruba	yor	latn	45996544	medlow	Niger-Congo
89	Haitian	hat	latn	43803264	medlow	Creole
90	Kinyarwanda	kin	latn	42016128	medlow	Niger-Congo
91	Samoan	smo	latn	41137664	medlow	Austronesian
92	Javanese	jav	latn	40730368	medlow	Austronesian
93	Norwegian Nynorsk	nno	latn	40680192	medlow	Indo-European
94	Lao	lao	laoo	40182528	medlow	Tai-Kadai
95	Nyanja	nya	latn	39635968	medlow	Niger-Congo
96	Sindhi	snd	arab	39586304	medlow	Indo-European
97	Southern Pashto	pbt	arab	39270656	medlow	Indo-European
98	Sundanese	sun	latn	39227648	medlow	Austronesian
99	Maori	mri	latn	39110528	medlow	Austronesian
100	Occitan	oci	latn	39094784	medlow	Indo-European
101	Plateau Malagasy	plt	latn	38467200	medlow	Austronesian
102	Pushto	pus	arab	37981184	medlow	Indo-European
103	Scottish Gaelic	gla	latn	37471488	medlow	Indo-European
104	Shona	sna	latn	37057152	medlow	Niger-Congo
105	Waray	war	latn	36727424	medlow	Austronesian
106	Zulu	zul	latn	36472960	medlow	Niger-Congo
107	Dari	prs	arab	36289920	medlow	Indo-European
108	Northern Uzbek	uzn	latn	35988736	medlow	Turkic
109	Uighur	uig	arab	35028992	medlow	Turkic
110	Assamese	asm	beng	34396032	medlow	Indo-European
111	Southern Sotho	sot	latn	34028544	medlow	Niger-Congo
112	Lushai	lus	latn	33796480	medlow	Sino-Tibetan
113	Standard Malay	zsm	latn	32638592	medlow	Austronesian
114	Xhosa	xho	latn	31847680	medlow	Niger-Congo
115	Sicilian	scn	latn	31407104	medlow	Indo-European
116	Lombard	lmo	latn	31299456	medlow	Indo-European
117	Eastern Yiddish	ydd	hebr	30456448	medlow	Indo-European
118	Egyptian Arabic	arz	arab	30198528	medlow	Afro-Asiatic
119	Limburgan	lim	latn	30182912	medlow	Indo-European
120	Odia	ory	orya	29186688	medlow	Indo-European
121	South Azerbaijani	azb	arab	29091584	medlow	Turkic
122	Ayacucho Quechua	quy	latn	29080448	medlow	Quechuan
123	West Central Oromo	gaz	latn	27978240	medlow	Afro-Asiatic
124	Halh Mongolian	khk	cyrl	27626624	medlow	Mongolic
125	Venetian	vec	latn	26978816	medlow	Indo-European
126	Banjar	bjn	latn	26552448	medlow	Austronesian
127	Gilaki	glk	arab	26084736	medlow	Indo-European
128	Ganda	lug	latn	25706752	medlow	Niger-Congo
129	Papiamento	pap	latn	24957568	medlow	Creole
130	Sanskrit	san	deva	24549760	medlow	Indo-European

131	Rundi	run	latn	24451072	medlow	Niger-Congo
132	Chinese	zho	hant	23736832	medlow	Sino-Tibetan
133	Achinese	ace	latn	23719936	medlow	Austronesian
134	Tswana	tsn	latn	23584384	medlow	Niger-Congo
135	Western Panjabi	pnb	arab	22000640	medlow	Indo-European
136	Twi	twi	latn	21262208	medlow	Atlantic-Congo
137	Iloko	ilo	latn	21032576	medlow	Austronesian
138	Chechen	che	cyrl	20793856	medlow	Nakh-Daghestanian
139	Tsonga	tso	latn	20281984	medlow	Niger-Congo
140	Yakut	sah	cyrl	19829248	medlow	Turkic
141	Western Frisian	fry	latn	19808384	medlow	Indo-European
142	Kurdish	kur	latn	19233152	medlow	Indo-European
143	Ewe	ewe	latn	18750848	medlow	Niger-Congo
144	Oriya	ori	orya	18473216	medlow	Indo-European
145	Latin	lat	latn	17430272	medlow	Indo-European
146	Chuvash	chv	cyrl	16924288	medlow	Turkic
147	Minangkabau	min	latn	16113024	medlow	Austronesian
148	Faroese	fao	latn	15750272	medlow	Indo-European
149	Breton	bre	latn	14796032	medlow	Indo-European
150	Yue Chinese	yue	hant	14777472	medlow	Sino-Tibetan
151	Pedi	nso	latn	14619264	medlow	Niger-Congo
152	Tosk Albanian	als	latn	14432000	medlow	Indo-European
153	Crimean Tatar	crh	latn	13975296	medlow	Turkic
154	Northern Kurdish	kmr	latn	13480832	medlow	Indo-European
155	Kabyle	kab	latn	13282688	medlow	Afro-Asiatic
156	Fon	fon	latn	13019904	medlow	Niger-Congo
157	Low German	nds	latn	12879488	medlow	Indo-European
158	Inuktitut	iku	cans	12683776	medlow	Eskimo-Aleut
159	Maithili	mai	deva	12227712	medlow	Indo-European
160	Lingala	lin	latn	12203136	medlow	Niger-Congo
161	Guarani	grn	latn	12139904	medlow	Tupian
162	Tibetan	bod	tibt	12052224	medlow	Sino-Tibetan
163	Pangasinan	pag	latn	11895296	medlow	Austronesian
164	Bemba	bem	latn	11693952	medlow	Niger-Congo
165	Wolof	wol	latn	11647872	medlow	Niger-Congo
166	Tumbuka	tum	latn	11176320	medlow	Atlantic-Congo
167	Luo	luo	latn	11028992	medlow	Eastern Sudanic
168	Malagasy	mlg	latn	10417152	low	Austronesian
169	Oromo	orm	latn	10022016	low	Afro-Asiatic
170	Dimli	diq	latn	9850112	low	Indo-European
171	Yiddish	yid	hebr	9727872	low	Indo-European
172	Tuvinian	tyv	cyrl	9700736	low	Turkic
173	Min Nan Chinese	nan	latn	9654656	low	Sino-Tibetan
174	Balinese	ban	latn	9067776	low	Austronesian
175	Fijian	fij	latn	8515328	low	Austronesian
176	Central Aymara	ayr	latn	8513792	low	Aymaran
177	Aragonese	arg	latn	8144384	low	Indo-European
178	Ligurian	lij	latn	7909120	low	Indo-European
179	Dhivehi	div	thaa	7748608	low	Indo-European
180	Luba-Lulua	lua	latn	7352192	low	Niger-Congo
181	Silesian	szl	latn	7311872	low	Indo-European
182	Nigerian Fulfulde	fuv	latn	6747136	low	Niger-Congo
183	Swiss German	gsw	latn	6581888	low	Indo-European
184	Swati	ssw	latn	6076160	low	Niger-Congo
185	Betawi	bew	cyrl	5948160	low	Creole
186	Friulian	fur	latn	5731584	low	Indo-European
187	Sardinian	srd	latn	5723904	low	Indo-European
188	Bavarian	bar	latn	5696512	low	Indo-European
189	Tok Pisin	tpi	latn	5505792	low	Creole
190	Umbundu	umb	latn	5479936	low	Niger-Congo
191	Nigerian Pidgin	pcm	latn	5292160	low	Creole
192	Eastern Mari	mhr	cyrl	5290752	low	Uralic
193	Ido	ido	latn	4775808	low	Constructed
194	Russia Buriat	bxr	cyrl	4556800	low	Mongolic
195	Bhojpuri	bho	deva	4365440	low	Indo-European
196	Bambara	bam	latn	4271232	low	Mande

197	Chokwe	cjk	latn	4177792	low	Atlantic-Congo
198	Southwestern Dinka	dik	latn	4137728	low	Nilotic
199	Dyula	dyu	latn	3980416	low	Mande
200	Mossi	mos	latn	3948544	low	Niger-Congo
201	Turkmen	tuk	latn	3940864	low	Turkic
202	Piemontese	pms	latn	3818368	low	Indo-European
203	Central Kanuri	knc	latn	3756288	low	Nilo-Saharan
204	Wu Chinese	wuu	hans	3689728	low	Sino-Tibetan
205	Kongo	kon	latn	3668224	low	Atlantic-Congo
206	Dargwa	dar	cyrl	3564800	low	Nakh-Daghestanian
207	Buginese	bug	latn	3539840	low	Austronesian
208	Kabuverdianu	kea	latn	3463936	low	Indo-European
209	Kabiyè	kbp	latn	3286272	low	Niger-Congo
210	Kimbundu	kmb	latn	3169536	low	Atlantic-Congo
211	Hawaiian	haw	latn	2996352	low	Austronesian
212	Sango	sag	latn	2924928	low	Niger-Congo
213	Mirandese	mwl	latn	2819584	low	Indo-European
214	Kachin	kac	latn	2732160	low	Sino-Tibetan
215	Ingush	inh	cyrl	2641408	low	Nakh-Daghestanian
216	Kikuyu	kik	latn	2636544	low	Niger-Congo
217	Romansh	roh	latn	2578304	low	Indo-European
218	Kaqchikel	cak	latn	2560256	low	Mayan
219	Kabardian	kbd	cyrl	2523264	low	Northwest Caucasian
220	Volapük	vol	latn	2522880	low	Constructed
221	Mandarin Chinese	cmn	hans	2511744	low	Sino-Tibetan
222	Kituba	mkw	cyrl	2431872	low	Creole
223	Magahi	mag	deva	2379776	low	Indo-European
224	Central Bikol	bcl	latn	2348672	low	Austronesian
225	Kashmiri	kas	deva	2302592	low	Indo-European
226	Cusco Quechua	quz	latn	2273280	low	Quechuan
227	Literary Chinese	lzh	hant	2267648	low	Sino-Tibetan
228	Walloon	wln	latn	2234880	low	Indo-European
229	Akan	aka	latn	2143360	low	Niger-Congo
230	Berber	ber	latn	2132352	low	Afro-Asiatic
231	Chhattisgarhi	hne	deva	2104576	low	Indo-European
232	Interlingua	ina	latn	2066816	low	Constructed
233	Upper Sorbian	hsb	latn	2062720	low	Indo-European
234	Latgalian	ltg	latn	2061952	low	Indo-European
235	Santali	sat	olck	1973888	low	Austro-Asiatic
236	Susu	sus	arab	1948160	low	Mande
237	Nuer	nus	latn	1941760	low	Eastern Sudanic
238	Vlaams	vls	latn	1928064	low	Indo-European
239	Quechua	que	latn	1901184	low	Quechuan
240	Udmurt	udm	cyrl	1857664	low	Uralic
241	Veps	vep	latn	1844736	low	Uralic
242	Avaric	ava	cyrl	1772288	low	Nakh-Daghestanian
243	Swahili	swh	latn	1768960	low	Niger-Congo
244	Lak	lbe	cyrl	1715328	low	Nakh-Daghestanian
245	Erzya	myv	cyrl	1714432	low	Uralic
246	Urdu	urd	deva	1697408	low	Indo-European
247	Ossetian	oss	cyrl	1697024	low	Indo-European
248	Uighur	uig	latn	1627648	low	Turkic
249	Lezghian	lez	cyrl	1625344	low	Nakh-Daghestanian
250	Goan Konkani	gom	deva	1604096	low	Indo-European
251	Shan	shn	mymr	1589248	low	Tai-Kadai
252	Serbian	srp	latn	1543424	low	Indo-European

Table 2: Languages included in our language modeling study.