Goldfish: Monolingual Language Models for 350 Languages

Tyler A. Chang[1], **Catherine Arnett**[2], **Zhuowen Tu**[1,3], **Benjamin K. Bergen**[1]
[1]Department of Cognitive Science, [2]Department of Linguistics,
[3]Department of Computer Science,
University of California San Diego

Abstract

For many low-resource languages, the only available language models are large multilingual models trained on many languages simultaneously. Despite state-of-the-art performance on reasoning tasks, we find that these models still struggle with basic grammatical text generation in many languages. First, large multilingual models perform worse than bigrams for many languages (e.g. 24% of languages in XGLM 4.5B; 43% in BLOOM 7.1B) using FLORES perplexity as an evaluation metric. Second, when we train small monolingual models with only 125M parameters on 1GB or less data for 350 languages, these small models outperform large multilingual models both in perplexity and on a massively multilingual grammaticality benchmark. To facilitate future work on low-resource language modeling, we release *Goldfish*, a suite of over 1,000 small monolingual language models trained comparably for 350 languages. These models represent the first publicly-available monolingual language models for 215 of the languages included.

Keywords: multilingual NLP, low-resource languages

🤗 Models and training data Training and evaluation code

1. Introduction

Language modeling research in low-resource languages often relies on large multilingual models trained on many languages simultaneously (Conneau et al., 2020b; Adelani et al., 2021b; Ebrahimi et al., 2022; Lin et al., 2022; Hangya et al., 2022; Imani et al., 2023). For many low-resource languages, a dedicated model optimized for that language does not exist. For example, in an analysis of publicly available models on Hugging Face, we find that 215 of the 350 languages in this paper did not have a monolingual text-generation model prior to the Goldfish models, and 47 of the languages did not have any text-generation models at all (Arnett and Chang, 2025).

This lack of dedicated models hinders comparability of results across models and languages (Bandarkar et al., 2024), and it contributes to model under-performance in low-resource languages (Wu and Dredze, 2020; Blasi et al., 2022). These barriers to research in low-resource languages are likely to exacerbate existing inequities across language communities in NLP research (Bender, 2011; Joshi et al., 2020a).

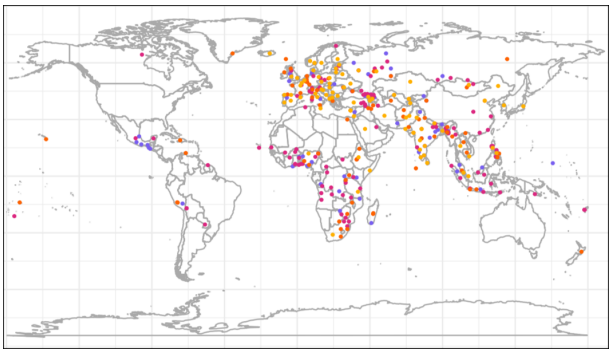
In particular, we find in this paper that large multilingual language models still struggle with basic next token prediction—a prerequisite for text generation—in low-resource languages. To establish a baseline for basic text generation performance in low-resource languages, we introduce **Goldfish**[1], a suite of over 1000 monolingual language models for 350 diverse languages.

The Goldfish reach lower perplexities than XGLM (Lin et al., 2022), BLOOM 7.1B (Scao et al., 2022), and MaLA-500 (Lin et al., 2024) on 98 out of 204 FLORES languages, despite each Goldfish model being over 10× smaller (§4.1). The Goldfish also outperform simple bigram models, which are surprisingly competitive with larger models for low-resource languages (e.g. lower perplexities than BLOOM 7.1B on 43% of its languages; §4.1). Finally, the Goldfish outperform large multilingual models on MultiBLiMP (Jumelet et al., 2025; §4.2), a massively multilingual grammaticality benchmark. However, despite better perplexities and grammaticality performance, the Goldfish perform at around chance, similar to other small multilingual models, on reasoning benchmarks (§4.2).

Thus, the primary contributions of this short paper are (1) to show that large multilingual models still suffer from poor next token prediction performance in low-resource languages, (2) to demonstrate that small monolingual models often exhibit lower perplexities and more grammatical next token predictions for such languages, and (3) to address the lack of available models in low-resource languages by releasing comparable monolingual models for 350 languages. Models, training data, and code are available at: https://huggingface.co/goldfish-models.

[1]This name refers to shared qualities between our models and goldfish (*Carassius auratus*); they are small, there are many of them, and they are known for their poor memories (perhaps inaccurately; Carey, 2024). If an acronym is desired, Goldfish can stand for **G**enerative aut**O**regressive **L**ow-resource mo**D**els **F**or l**I**mited-compute **S**ystem **H**ardware.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3750–3781
11–16 May 2026.



Data size	Model output
5MB	`Goldfish are a few years of the most of the most of the most...`
10MB	`Goldfish are a great way to the best way to the best way...`
100MB	`Goldfish are a great way to get your fish in the wild.`
1GB	`Goldfish are a species of fish that are found in the sea.`

Figure 1: Left: Map of the 350 languages for which Goldfish models are available, using coordinates from Glottolog (Hammarström et al., 2023). Color indicates the largest dataset size for that language. Right: Sample model outputs completing the prompt `"Goldfish are"` for the `eng_latn` (English) model for each dataset size, using sampling temperature zero. Grammatical text generation begins to emerge in the 100MB-dataset model (available for 166 languages), but the lower-resource models still achieve better perplexities than previous models for many low-resource languages (§4.1).

2. Related Work

Low resource language modeling often leverages multilingual pretraining, where a model is trained on multiple languages simultaneously (Pires et al., 2019; Conneau et al., 2020b). Indeed, this can improve low-resource performance, particularly when models have sufficient capacity and the multilingual data is from related or typologically similar languages (Kakwani et al., 2020; Ogueji et al., 2021; Chang et al., 2024a). However, monolingual models have still been shown to achieve better performance than multilingual models for many languages (e.g. Martin et al., 2020; Pyysalo et al., 2021; Gutiérrez-Fandiño et al., 2021; Luukkonen et al., 2023). Thus, it appears that existing multilingual language models are still limited by model capacity or limited data in low-resource languages (Conneau et al., 2020b; Chang et al., 2024a).

Notably, the training datasets for massively multilingual models are often heavily skewed towards high-resource languages. For example, XGLM 4.5B is trained on over 7000× more Norwegian (71GB; 5.4M native speakers) than Quechua (0.01GB; 7.3M native speakers; Lin et al., 2022; Ethnologue, 2024). In a more extreme case, BLOOM is trained on only 0.07MB of Akan (8.1M native speakers) out of 1.61TB total (4e-6% of the pretraining dataset; Scao et al., 2022). These extremely small quantities of low-resource language data often do not leverage recent efforts to compile text data in low-resource languages (Costa-jussà et al., 2022; Imani et al., 2023; Kudugunta et al., 2023), and the data imbalances are likely to severely hinder performance in low-resource languages. Indeed, we find that these models have worse perplexities than simple bigram models for many languages (§4.1). Unfortunately, comparable monolingual language models across many diverse languages have yet to be studied or released.

3. Models and Datasets

To compare to larger multilingual models, we introduce the Goldfish models, a suite of 1154 monolingual Transformer language models pretrained for 350 languages. The largest model for each language is 125M parameters. We train models on 5MB, 10MB, 100MB, and 1GB of text when available after byte premium scaling (Arnett et al., 2024). Figure 1 shows a geographic map of the 350 languages, with coordinates from Glottolog (Hammarström et al., 2023), along with sample outputs from the English model for each dataset size.

3.1. Training Datasets

We merge the massively multilingual text datasets compiled in Chang et al. (2024a), Glot500 (Imani et al., 2023), and MADLAD-400 (Kudugunta et al., 2023) per language. We deduplicate repeated sequences of 100 UTF-8 bytes and drop languages with only Bible data. Full dataset details are in §A.1. To facilitate fair evaluations, we hold out FLORES-200 and AmericasNLI from all datasets (Costa-jussà et al., 2022; Ebrahimi et al., 2022). We conduct a contamination analysis to determine whether FLORES is contaminated in our datasets and find that for 98% of languages, less than 10 out of 2000 FLORES sequences appear in the training dataset at all (§A.2).

To sample pretraining datasets of the desired sizes in a language L, we first use the Byte Premium Tool (Arnett et al., 2024) to estimate the

byte premium for L, the number of UTF-8 bytes required to encode comparable text in L relative to `eng_latn` (English). For example, `khm_khmr` (Khmer) has byte premium 3.91, meaning that it uses approximately $3.91\times$ as many UTF-8 bytes as English to encode content-matched text. We divide each dataset size by the estimated byte premium for the corresponding language, thus measuring all datasets in units of "equivalent" English text bytes. We sample datasets to train monolingual language models on 5MB, 10MB, 100MB, and 1GB when available after byte premium scaling.[2] These are equivalent to roughly 1M, 2M, 20M, and 200M tokens of English text respectively; including 10 epochs of repetition, the 1GB-dataset models are trained on the equivalent of roughly 2B English tokens. When a 1GB dataset is not available for a language after byte premium scaling, we include a **full** model (267 languages) trained on the entire dataset in that language, for use cases that seek to maximize performance in a specific low-resource language.

As we cap our dataset sizes at 1GB per language, our models are trained on much less data than is otherwise available for high-resource languages. Still, when we compare our dataset sizes to FineWeb2 (Penedo et al., 2025) for the languages covered by both Goldfish and FineWeb2, 106 of those languages have more data in our dataset than in FineWeb2. We attribute this partially to our approach to data collection, which involves seeking out language-specific resources rather than relying solely on language identification models, which can be unreliable for web data in low-resource languages (Ortiz Suarez et al., 2026). We release our training datasets at: `https://huggingface.co/datasets/goldfish-models/fish-food`.

3.2. Architectures and Pretraining

We train monolingual language models for five dataset sizes when available after byte premium scaling: **5MB**, **10MB**, **100MB**, **1GB**, and **full**. The full dataset size (including all available data) is only included if a 1GB dataset is not available for a language. In total, the Goldfish include 350 5MB-dataset models, 288 10MB-dataset models, 166 100MB-dataset models, 83 1GB-dataset models, and 267 full-dataset models (1154 models total).

For each language and each dataset size, we pretrain an autoregressive GPT-2 Transformer language model from scratch (Radford et al., 2019). For the 1GB, 100MB, and full dataset sizes, we use the 125M-parameter architecture equivalent to GPT-1 (Radford et al., 2018), which has a similar parameter count to BERT-base and RoBERTa (Devlin et al., 2019; Liu et al., 2019). Because larger models do not appear to outperform smaller models for very small datasets (Chang et al., 2024a), we use the small model size (39M parameters) from Turc et al. (2019) for the 10MB and 5MB dataset sizes. Full hyperparameters are reported in §A.3.

Each model has a custom monolingual tokenizer, which is trained with a vocabulary size of 50K (Liu et al., 2019) on the same dataset size as their corresponding model (including byte premium scaling). We use Unigram tokenizers trained with the SentencePiece implementation (Kudo and Richardson, 2018). Training text is randomly sampled from the dataset for the desired language.[3] After tokenizer training, we tokenize each training dataset, concatenating text lines such that each sequence contains exactly 512 tokens. We run tokenization before shuffling and sampling to the desired dataset sizes, so our sequences of 512 tokens preserve contiguous text where possible, although several of our source corpora only exist in shuffled form. Finally, we sample our tokenized datasets to 5MB, 10MB, 100MB, and 1GB after byte premium scaling.[4]

We train each model for 10 epochs of the training data; multiple epochs of pretraining is beneficial in data-constrained scenarios (Muennighoff et al., 2023), but pretraining on more than 10 epochs often leads to overfitting (increases in eval loss) in the 5MB scenarios. All language model pretraining runs together take a total of 1.65×10^{20} FLOPs. This is less than $1/1900\times$ the computation used to train the original 175B-parameter GPT-3 model (Brown et al., 2020a; 3.14×10^{23} FLOPs). Further pretraining details are reported in §A.3.

3.3. Training Bigram Baselines

Because there do not exist other models for many of these languages, we also train bigram baseline models on the same training data and with the same tokenizer for each language. Each bigram model computes the probability of each token w_i as $P(w_i|w_{i-1})$, computed based on raw bigram counts in the tokenized Goldfish dataset. When a bigram is not observed in the dataset, we use backoff to unigram probability with a penalty multiplier of $\lambda = 0.40$ (i.e. "stupid backoff"; Brants et al., 2007). We do not consider n-grams for $n > 2$ because those n-grams often resort to backoff and are therefore much more sensitive to the backoff penalty term λ.

[2]The languages with 5MB-dataset models are a subset of the languages with 10MB-dataset models, and similarly for the 100MB and 1GB dataset sizes.

[3]To avoid memory errors, we limit tokenizer training text to 100MB after byte premium scaling.

[4]When de-tokenized, the tokenized datasets are slightly smaller than the original text datasets, because the tokenizer truncates lines to create 512-token sequences. Reported dataset sizes account for truncation.

Goldfish data size	# Langs	Goldfish	Bigrams	XGLM 4.5B	MaLA-500 10B
1000MB	73	**76.9**	112.3	78.6	84.7
100MB	22	**102.7**	132.6	143.9	121.7
10MB, 5MB	5	**130.5**	148.3	183.1	135.0

Table 1: Mean FLORES perplexity (↓) for the 100 languages in XGLM 4.5B, MaLA-500, and FLORES, separated by maximum Goldfish dataset size. The Goldfish languages are a strict superset of these languages.

	Bigrams	XGLM 4.5B	XGLM 7.5B	BLOOM 7.1B	MaLA-500 10B
Bigrams		24 / 102	0 / 30	20 / 46	11 / 175
Goldfish (ours)	**202** / 202	**60** / 102	2 / 30	**32** / 46	**111** / 175

Table 2: FLORES perplexity win rates for each row vs. column model. For example, Goldfish reach lower log-perplexities than MaLA-500 for 111/175 (63%) of FLORES languages in both Goldfish and MaLA-500.

4. Evaluations

4.1. FLORES Log-Perplexity

We first evaluate our models on FLORES-200 log-perplexity (Costa-jussà et al., 2022) (equivalently, negative log-likelihood; Lin et al., 2024). To avoid confounds from different tokenizers across models, we compute log-perplexities at the sequence level. Specifically, regardless of its tokenization, a language model $\mathcal{M}$ assigns some probability $P_{\mathcal{M}}(s)$ to each sequence s in FLORES. In most cases, s is a single sentence. For fair comparison with multilingual models that need to determine the input language during the early parts of a sequence, we compute log-perplexity of the second half s_1 of each sequence given the first half s_0. We then compute the mean over sequences:

$$\text{LogPPL}_{\mathcal{M}} = \text{mean}_s\Big(-\log(P_{\mathcal{M}}(s_1|s_0))\Big) \quad (1)$$

A lower log-perplexity indicates better performance, where $\mathcal{M}$ assigns higher probabilities to ground truth text (FLORES sequences). While imperfect, perplexity does not require annotated text data, it is predictive of performance on a variety of downstream tasks (Xia et al., 2023), and it has been used to measure language model quality in previous work (Kaplan et al., 2020; Hoffmann et al., 2022; Lin et al., 2024).

We compare the Goldfish with XGLM 4.5B (Lin et al., 2022; 134 languages), XGLM 7.5B (30 languages), BLOOM 7.1B (Scao et al., 2022; 46 languages), and MaLA-500 10B (Lin et al., 2024; 534 languages). We also compare to our bigram baselines, which were trained on the same datasets as the Goldfish models. In all cases, we use the Goldfish model trained on the maximum amount of data in a language (maximum 1GB).

The Goldfish reach lower log-perplexities than all four comparison models on 98 of the 204 FLORES languages. On average, the Goldfish reach 13% lower perplexities than XGLM 4.5B, and 11% lower than MaLA-500 10B (Table 1). To ensure that these results are not driven by a small subset of specific languages, in Table 2 we also report the pairwise "win" rates for Goldfish and bigrams vs. all four comparison models, for the set of FLORES languages shared between each pair. The Goldfish models have a perplexity win rate above 50% against all comparison models except XGLM 7.5B, which considers only 30 fairly high-resource languages (Lin et al., 2022). Notably, the *bigram* models also reach lower perplexities than large multilingual models for a nontrivial number of languages: 24% of languages in XGLM 4.5B and 43% of languages in BLOOM 7.1B. Still, the bigrams have worse perplexities than Goldfish for all languages. Perplexities for individual languages and models are available in our GitHub.

4.2. Downstream Tasks

Next, we evaluate linguistic knowledge (grammaticality) with MultiBLiMP (Jumelet et al., 2025) as implemented in the LM Evaluation Harness (Gao et al., 2023); this covers 74 of the Goldfish languages. We compare against popular small multilingual models: BLOOM 560M (Scao et al., 2022), XGLM (564M and 1.7B; Lin et al., 2021), Gemma 3 (270M and 1B base models; Gemma Team, 2025), and Llama 3.2 (1B base model; Meta AI, 2024). Results are in Table 3. The Goldfish models have higher average accuracy than any of the multilingual models, and they have the highest accuracy of all models tested for 25 of the 74 languages. This result highlights the benefits of small monolingual models, especially for languages which only account for a small portion of the training data in multilingual models.

We also evaluate the same models on three popular multilingual reasoning benchmarks: Belebele (121 languages, reading comprehension; Ban-

Dataset / Model	Chance	Goldfish 124M	Gemma 3 270M	BLOOM 560M	XGLM 564M	Gemma 3 1B	LLaMA 3 1B	XGLM 1.7B
MultiBLiMP (avg)	50.0	**78.8**	72.3	64.2	66.6	78.0	77.4	62.7
Belebele (avg)	25.0	28.2	22.7	28.9	**29.0**	22.7	28.9	28.2
XCOPA (avg)	50.0	55.1	54.6	54.4	55.2	**59.7**	55.9	57.6
XStoryCloze (avg)	50.0	52.3	52.7	52.6	53.1	**58.5**	55.1	56.2

Table 3: Average performance on downstream benchmarks for all models tested.

darkar et al., 2024), XCOPA (11 languages, commonsense reasoning; Ponti et al., 2020), and XStoryCloze (10 languages, story commonsense reasoning; Lin et al., 2022). All models are evaluated zero shot with log-probability solution ranking, with no fine-tuning or instruction tuning. Unfortunately, all models perform quite poorly (close to chance accuracy; Table 3), suggesting that available multilingual reasoning evaluations are not appropriate for pretrained-only models of this scale.

5. Discussion

Our results demonstrate that large multilingual language models still struggle with basic grammatical text generation for many languages (e.g. often worse perplexities than bigrams). The Goldfish (small monolingual models) exhibit lower perplexities and more grammatical next token predictions for many low-resource languages. These results make the Goldfish suitable baselines for basic grammatical text generation in diverse languages, motivating future work developing low-resource language models. Furthermore, the Goldfish are accessible to labs with limited compute budgets, they are trained on a roughly human-like amount of data (Warstadt et al., 2023), and they are trained to be maximally comparable across languages. Future work may investigate precisely when larger-scale multilingual pretraining provides benefits to lower-resource languages; for example, it may be that abstract reasoning patterns and heuristics are often more language-agnostic than grammatical text generation, and thus larger-scale multilingual pretraining primarily benefits the former.

6. Conclusion

In this paper, we pretrain and release Goldfish, a suite of over 1000 monolingual language models for 350 languages. For the majority of these languages, the Goldfish represent the first monolingual model dedicated for that language. The Goldfish achieve perplexities that are competitive with, and on average lower than, state-of-the-art multilingual language models across languages. However, similar to multilingual models of the same scale, the Goldfish still struggle with reasoning tasks. This suggests that smaller monolingual models may better represent linguistic knowledge of the target language, but they do not perform well on more complex tasks at this scale. We publicly release the Goldfish models to promote future research pushing the limits of language model capabilities in low-resource languages.

Limitations

Comparability and availability. In order to include as many low-resource languages as possible, the Goldfish models are trained on corpora compiled from a wide variety sources (§A.1). Still, 5MB of text (roughly 1M tokens) is not publicly available for many of the world's languages. Even where text is available, corpora for different languages vary significantly both in cleanliness and domain coverage (e.g. news vs. social media vs. books). Thus, while we release models trained on comparable quantities of text in different languages (including accounting for byte premiums; Arnett et al., 2024; §3.1), the models are not perfectly comparable across languages. In fact, it is likely that such perfect comparability is impossible given the diversity of the world's languages, cultures, and language use. Even directly translated datasets are not perfectly comparable across languages (Jill Levine and Lateef-Jan, 2018). Thus, the Goldfish models aim to maximize model and dataset comparability across languages while still covering a wide variety of languages.

Monolinguality. By design, all of the Goldfish models are monolingual. For low-resource languages, training on closely related languages would likely improve performance (Conneau et al., 2020b; Chang et al., 2024a). However, adding multilingual data introduces concerns such as the choice of added languages (some languages have more closely related languages in our dataset than others), quantities of added data, and model capacity limitations. To maximize comparability across languages and to allow the models to serve as clearly-defined baselines, we train all Goldfish models monolingually. Of course, language-annotated text datasets inevitably contain mislabeled text, particularly for similar languages (Caswell et al., 2020;

Blevins and Zettlemoyer, 2022; Kreutzer et al., 2022). Thus, we cannot guarantee that our models are entirely free from cross-language contamination, although they are monolingual to the best ability of current language identification models.

Model and dataset sizes. Because the Goldfish are focused on low-resource languages, we restrict all models to 1GB of training text (after byte premium scaling; Arnett et al., 2024). For the majority of the world's languages, 1GB is sufficient to include all publicly available text data in the language. At these small dataset sizes, larger models do not appear to provide significant benefit over smaller models (Kaplan et al., 2020; Hoffmann et al., 2022; Chang et al., 2024a). Thus, the largest Goldfish model that we train for each language has 125M parameters and is trained on a maximum of 1GB of text. This is the same model size as GPT-1 (Radford et al., 2018) or BERT (Devlin et al., 2019), and the 1GB dataset size is approximately 20% of the dataset size of GPT-1 (Radford et al., 2018).

Downstream tasks. We evaluate the Goldfish models on FLORES log-perplexity (§4.1) and four multilingual benchmarks, including three reasoning benchmarks (§4.2). Perplexity is the only evaluation available for autoregressive language models in many languages before instruction-tuning and RLHF, but it has significant limitations. Perplexity is not necessarily predictive of more specific capabilities (Hu et al., 2020a; Levy et al., 2024), although it still provides reasonable signal for model performance (Xia et al., 2023). On the other hand, reasoning benchmarks require annotated datasets and thus often cover fewer languages. Belebele (121 non-English languages; Bandarkar et al., 2024) is an exception, but even state-of-the-art models perform quite poorly on Belebele without instruction-tuning or few-shot prompting (§4.2). Thus, our evaluations of model reasoning are not conclusive; we may primarily be measuring heuristics that allow the models to perform only somewhat above chance (arguably, this might still be considered a basic form of "reasoning").

Outside of reasoning benchmarks, we also evaluate on MultiBLiMP (Jumelet et al., 2025), which evaluates linguistic knowledge. MultiBLiMP significantly expands the language coverage of multilingual grammaticality benchmarks; however, MultiBLiMP evaluates a very narrow aspect of grammatical knowledge (two types of subject-verb agreement). We hope that tractable evaluation datasets with broad language coverage will become increasingly available in the future to enable more informative evaluation of models in a broader range of languages.

Risks and dataset licensing. Trained on a maximum of 1GB of text each, the Goldfish models have very limited capabilities relative to modern language models in high-resource languages. The Goldfish are trained on publicly-released corpora used in previous NLP research (§A.1), but we cannot guarantee that the data is free from offensive content or personally identifying information. Our models are small, which reduces the likelihood that they will regurgitate memorized text (Carlini et al., 2023). As far as we are aware, we do not include any datasets that prohibit use for language model training. We report all included datasets in §A.1. We will remove models for affected languages if contacted by dataset owners.

7. Bibliographical 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.

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

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

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. SERENGETI: Massively multilingual language models for Africa. *arXiv preprint arXiv:2212.10785*.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin

Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

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

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. 2021b. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

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

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

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.

AI FOR THAI. 2023. Ai for thai lotuscorpus. Dataset.

AI4Bharat. 2023. AI4Bharat. Dataset.

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.

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

American Psychological Association. 1983. *Publications Manual*. American Psychological Association, Washington, DC.

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

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L_1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796.

Anuvaad. 2023. Anuvaad project. Dataset.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale english language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068.

Catherine Arnett and Tyler Chang. 2025. An Analysis of Multilingual Models on Hugging Face. Hugging Face blog post.

Catherine Arnett, Tyler A. Chang, and Benjamin Bergen. 2024. A bit of a problem: Measurement disparities in dataset sizes across languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 1–9, Torino, Italia. ELRA and ICCL.

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

Autshumato. 2023. Autshumato. Dataset.

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

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele benchmark: A parallel reading comprehension dataset in 122 language variants. In *Annual Meeting of the Association for Computational Linguistics*.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Marta Bañón, Miquel Esplà-Gomis, Mikel L. Forcada, Cristian García-Romero, Taja Kuzman, Nikola Ljubesic, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, Tobias van der Werff, and Jaume Zaragoza. 2022. Macocu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation, EAMT 2022, Ghent, Belgium, June 1-3, 2022*, pages 301–302. European Association for Machine Translation.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Emily Bender. 2019. The# benderrule: On naming the languages we study and why it matters. *The Gradient*, 14(1).

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.

Workshop BigScience, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar

Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter,

Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: a 176b-parameter open-access multilingual language model.

Jan Christian Blaise Cruz and Charibeth Cheng. 2020. Establishing baselines for text classification in low-resource languages. *arXiv e-prints*, pages arXiv–2005.

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.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. *arXiv*.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. 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.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, 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. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

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

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

José Ramom Pichel Campos, Pablo Gamallo Otero, and Iñaki Alegria Loinaz. 2020. Measuring diachronic language distance using perplexity: Application to english, portuguese, and spanish. *Natural Language Engineering*, 26(4):433–454.

Lily Carey. 2024. Goldfish may have a longer memory span than just three seconds. *Discover Magazine*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *International Conference on Learning Representations*.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2024a. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. 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.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2024b. Characterizing learning curves during language model pre-training: Learning, forgetting, and stability. *Transactions of the Association for Computational Linguistics*.

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

Ethan A Chi, John Hewitt, and Christopher D Manning. 2020. Finding universal grammatical relations in multilingual bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. 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 3418–3430, Online. Association for Computational Linguistics.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).

Clarin. 2023. Clarin.si. Dataset.

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

Common Crawl. 2022. Common crawl. Dataset.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020a. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

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

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020c. 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.

James W. Cooley and John W. Tukey. 1965. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation*, 19(90):297–301.

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*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya Expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv*.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

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.

Jonathan Dunn. 2020. Mapping languages: the corpus of global language use. *Lang. Resour. Evaluation*, 54(4):999–1018.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2022. Ethnologue: Languages of the world. twenty-fifth edition.

eBible. 2023. eBible. Dataset.

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.

Mahmoud El-Haj. 2020. Habibi - a multi dialect multi national Arabic song lyrics corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1318–1326, Marseille, France. European Language Resources Association.

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

Obeida ElJundi, Wissam Antoun, Nour El Droubi, Hazem Hajj, Wassim El-Hajj, and Khaled Shaban. 2019. hULMonA: The universal language model in Arabic. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 68–77, Florence, Italy. 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.

Ethnologue. 2024. *Ethnologue, Languages of the World*. SIL International.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

FFR Dataset. 2023. Fon and french dataset. Dataset.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*.

Negar Foroutan, Mohammadreza Banaei, Remi Lebret, Antoine Bosselut, and Karl Aberer. 2022. Discovering language-neutral sub-networks in multilingual language models. *arXiv preprint arXiv:2205.12672*.

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)*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. Language model evaluation harness: A framework for few-shot language model evaluation.

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.

Google DeepMind Gemma Team. 2025. Gemma 3 technical report. *arXiv*.

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).

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2021. Experiments on a Guarani corpus of news and social media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 153–158, Online. Association for Computational Linguistics.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2022. Can we use word embeddings for enhancing Guarani-Spanish machine translation? In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 127–132, Dublin, Ireland. Association for Computational Linguistics.

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

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33.

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.

Dan Gusfield. 1997. *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. MarIA: Spanish language models. *arXiv*.

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

Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovickỳ, Constantin A Rothkopf, Alexander Fraser, and Kristian Kersting. 2022. Speaking multiple languages affects

the moral bias of language models. *arXiv preprint arXiv:2211.07733*.

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.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Samin Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-Sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4693–4703. Association for Computational Linguistics.

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.

HornMT. 2023. Machine translation benchmark dataset for languages in the horn of africa. Dataset.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020a. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompt-based methods may underestimate large language models' linguistic generalizations. *arXiv preprint arXiv:2305.13264*.

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

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020c. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

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, Toronto, Canada. Association for Computational Linguistics.

Ayyoob ImaniGooghari, Silvia Severini, Masoud Jalili Sabet, François Yvon, and Hinrich Schütze. 2022. Graph-based multilingual label propagation for low-resource part-of-speech tagging. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1577–1589, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Suzanne Jill Levine and Katie Lateef-Jan. 2018. *Untranslatability Goes Global*. Routledge.

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.

Alexander Jones, William Yang Wang, and Kyle Mahowald. 2021. A massively multilingual analysis of cross-linguality in shared embedding space. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5833–5847.

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. *arXiv preprint arXiv:1912.03457*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020a. 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.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020b. 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.

Jaap Jumelet, Leonie Weissweiler, and Arianna Bisazza. 2025. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv*. URL removed because this paper includes un-anonymized links to the current paper. This paper is published by authors unaffiliated with the authors of the current paper.

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*.

Slava Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):400–401.

Philipp Koehn. 2023. Statistical and neural machine translation. Dataset.

Fajri Koto and Ikhwan Koto. 2020. Towards computational linguistics in Minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 138–148, Hanoi, Vietnam. Association for Computational Linguistics.

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

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

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. *arXiv*.

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

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The BigScience ROOTS Corpus: A 1.6 TB Composite Multilingual Dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8608–8621. Association for Computational Linguistics.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pretrained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674.

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

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.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

LINDAT. 2023. Lindat/clariah-cz repository. Dataset.

Lingala Songs. 2023. Lingala song lyrics. Dataset.

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.

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

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does RoBERTa know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. 2023. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726, Singapore. Association for Computational Linguistics.

LyricsTranslate. 2023. Lyricstranslate. Dataset.

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

Rooweither Mabuya, Jade Abbott, and Vukosi Marivate. 2023. Umsuka isizuluparallel corpus. Dataset.

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.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*.

Martin Majliš. 2011. W2C – web to corpus – corpora. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL),

Faculty of Mathematics and Physics, Charles University.

Dan Malkin, Tomasz Limisiewicz, and Gabriel Stanovsky. 2022. A balanced data approach for evaluating cross-lingual transfer: Mapping the linguistic blood bank. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4903–4915, Seattle, United States. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Masakhane. 2023. Masakhane: A living collection of NLP projects for Africans, by Africans. Dataset.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*, pages 3158–3163. European Language Resources Association (ELRA).

Meta. 2024. The Llama 3 herd of models. *arXiv*.

Meta AI. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

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

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

Makoto Morishita, Jun Suzuki, and Masaaki Nagata. 2020. JParaCrawl: A large scale web-based English-Japanese parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3603–3609, Marseille, France. European Language Resources Association.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems*.

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

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.

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

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

Nart. 2023. Abkhaz text. Dataset.

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

Patrick Niyongabo. 2023. An english-kinyarwanda statistical machine translation (SMT) model. Dataset.

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 Ortiz Suarez, Laurie Burchell, Catherine Arnett, Rafael Mosquera-Gómez, Sara Hincapie-Monsalve, Thom Vaughan, Damian Stewart, Malte Ostendorff, Idris Abdulmumin, Vukosi Marivate, et al. 2026. Commonlid: Re-evaluating state-of-the-art language identification performance on web data. *arXiv preprint arXiv:2601.18026*.

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.

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

Sungjoon Park, Sungdong Kim, Jihyung Moon, Won Ik Cho, Kyunghyun Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, et al. 2021. Klue: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*. Advances in Neural Information Processing Systems.

Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, and Chris Callison-Burch. 2022. Bidirectional language models are also few-shot learners. *arXiv preprint arXiv:2209.14500*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all — adapting pre-training data processing to every language. In *Second Conference on Language Modeling*.

Pepperidge Farm. 2024. Goldfish® Crackers. [Online; accessed 9-July-2024].

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. Unks everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203.

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.

Marco Polignano, Valerio Basile, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. Alberto: Modeling italian social media language with bert. *IJCoL. Italian Journal of Computational Linguistics*, 5(5-2):11–31.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

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*.

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

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.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Machel Reid and Mikel Artetxe. 2022. On the role of parallel data in cross-lingual transfer learning. *arXiv preprint arXiv:2212.10173*.

Roberts Rozis and Raivis Skadiņš. 2017. Tilde MODEL - multilingual open data for EU languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, Gothenburg, Sweden. Association for Computational Linguistics.

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.

SADiLaR. 2023a. Mburisano covid-19 multilingual corpus. Dataset.

SADiLaR. 2023b. South african centre for digital language resources, nchlt corpus. Dataset.

Hassan Sajjad, Ahmed Abdelali, Nadir Durrani, and Fahim Dalvi. 2020. AraBench: Benchmarking dialectal Arabic-English machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5094–5107, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Elizabeth-Jane Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Franccois 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*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Silvia Severini, Ayyoob Imani, Philipp Dufter, and Hinrich Schütze. 2022. Towards a broad coverage named entity resource: A data-efficient approach for many diverse languages. *arXiv preprint arXiv:2201.12219*.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. 2022. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *arXiv preprint arXiv:2201.03110*.

Anil Kumar Singh. 2008a. Named entity recognition for south and south East Asian languages: Taking stock. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*.

Anil Kumar Singh. 2008b. Natural language processing for less privileged languages: Where do we come from? where are we going? In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Stanford. 2023. Stanford nlp group datasets. Dataset.

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.

Team Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova,

Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. 2018. Parallel corpora for bi-directional statistical machine translation for seven Ethiopian language pairs. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 83–90, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

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*.

Iulia Turc, Kenton Lee, Jacob Eisenstein, Ming-Wei Chang, and Kristina Toutanova. 2021. Revisiting the primacy of english in zero-shot cross-lingual transfer. *CoRR*, abs/2106.16171.

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.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze. 2023. NLNDE at semeval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. *CoRR*, abs/2305.00090.

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.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Wikimedia. 2023. Wikimedia dumps.

Wikipedia. 2024. Wikipedia.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

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.

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

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.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Jian Yang, Shuming Ma, Dongdong Zhang, Shuangzhi Wu, Zhoujun Li, and Ming Zhou. 2020. Alternating language modeling for cross-lingual pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9386–9393.

Lyudmila Zaydelman, Irina Krylova, and Boris Orekhov. 2016. The technology of web-texts collection of Russian minor languages. In *Proceed-*

ings of the International Scientific Conference CPT2015, pages 179–181.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Nuria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022. Introducing qubert: A large monolingual corpus and bert model for southern quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13.

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.

Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, and Hinrich Schütze. 2020. Masking as an efficient alternative to finetuning for pretrained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2226–2241.

Zhong Zhou and Alex Waibel. 2021. Family of origin and family of choice: Massively parallel lexiconized iterative pretraining for severely low resource machine translation. *arXiv preprint arXiv:2104.05848*.

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. Appendix

A.1. Training Dataset Details

Data sources. As described in §3.1, we merge the text datasets compiled in Chang et al. (2024a), Glot500 (Imani et al., 2023), and MADLAD-400 (clean split; Kudugunta et al., 2023). These datasets include popular multilingual corpora such as OSCAR (Ortiz Suárez et al., 2019; Abadji et al., 2021), Wikipedia (Wikipedia, 2024), No Language Left Behind (Costa-jussà et al., 2022), and others. Together, these datasets take advantage of both automatically crawled datasets with automated language identification and targeted datasets manually annotated for specific low-resource languages. All included datasets are publicly available; see Limitations for licensing concerns. Comprehensively, the Goldfish dataset includes:

- Chang et al. (2024a):
  OSCAR (Ortiz Suárez et al., 2019; Abadji et al., 2021), Wikipedia (Wikipedia, 2024), No Language Left Behind (Costa-jussà et al., 2022), Leipzig Corpora Collection (Goldhahn et al., 2012), eBible translations (eBible, 2023), Tatoeba (Tiedemann, 2012, 2020), AfriBERTa (Ogueji et al., 2021), NusaX (Winata et al., 2023), AmericasNLP (Mager et al., 2021), Nunavut Hansard Inuktitut–English Parallel Corpus (Joanis et al., 2020), Cherokee-English ChrEn dataset (Zhang et al., 2020), Cherokee Corpus (Cherokee Corpus, 2023), Cree Corpus (Teodorescu et al., 2022), Languages of Russia (Zaydelman et al., 2016), Evenki Life newspaper (Zueva et al., 2020), transcribed Fula Speech Corpora (Cawoylel, 2023), IsiXhosa (Podile and Eiselen, 2016), Ewe Language Corpus (Gbedevi Akouyo et al., 2021), Makerere Luganda Corpora (Mukiibi et al., 2022), CMU Haitian Creole dataset (CMU, 2010), Tigrinya Language Modeling Dataset (Gaim et al., 2021), and Ulukau (Ulukau, 2023).
- Glot500 (Imani et al., 2023):
  AI4Bharat (AI4Bharat, 2023), AI FOR THAI LotusCorpus (AI FOR THAI, 2023), Arabic Dialects Dataset (El-Haj et al., 2018), AfriBERTa (Ogueji et al., 2021), AfroMAFT (Adelani et al., 2022; Xue et al., 2021), Anuvaad (Anuvaad, 2023), AraBench (Sajjad et al., 2020), Autshumato (Autshumato, 2023) Bloom Library (Leong et al., 2022), CC100 (Conneau et al., 2020b), CCNet (Wenzek et al., 2020), CMU Haitian Creole (CMU, 2010), SADiLaR NCHLT corpus (SADiLaR, 2023b), Clarin (Clarin, 2023), DART (Alsarsour et al., 2018), Earthlings (Dunn, 2020), FFR Dataset (FFR Dataset, 2023), GiossaMedia (Góngora et al., 2022, 2021), Glosses (Camacho-Collados et al., 2016), Habibi (El-Haj, 2020), HinDialect (Bafna, 2022), HornMT (HornMT, 2023), IITB (Kunchukuttan et al., 2018), IndicNLP (Nakazawa et al., 2021), Indiccorp (Kakwani et al., 2020), isiZulu (Mabuya et al., 2023), JParaCrawl (Morishita et al., 2020), kinyarwandaSMT (Niyongabo, 2023), LeipzigData (Goldhahn et al., 2012), LINDAT (LINDAT, 2023), Lingala Song Lyrics (Lingala Songs, 2023), LyricsTranslate (LyricsTranslate, 2023), mC4 (Raffel et al., 2020), MTData (Gowda et al., 2021), MaCoCu (Bañón et al., 2022), Makerere MT Corpus (Mukiibi et al., 2022), Masakhane Community (Masakhane, 2023), Mburisano Covid Corpus (SADiLaR, 2023a), Menyo20K (Adelani et al., 2021a), Minangkabau corpora (Koto and Koto, 2020), MoT (Palen-Michel et al., 2022), NLLB seed (Costa-jussà et al., 2022), Nart Abkhaz text (Nart, 2023), OPUS (Tiedemann, 2012), OSCAR (Ortiz Suárez et al., 2019), ParaCrawl (Bañón et al., 2020), Parallel Corpora for Ethiopian Languages (Teferra Abate et al., 2018), Phontron (Neubig, 2011), QADI (Abdelali et al., 2021), Quechua-IIC (Zevallos et al., 2022), SLI GalWeb.1.0 (Agerri et al., 2018), Shami (Abu Kwaik et al., 2018), Stanford NLP (Stanford, 2023), StatMT (Koehn, 2023), TICO (Anastasopoulos et al., 2020), TIL (Mirzakhalov et al., 2021), Tatoeba (Tiedemann, 2020), TeDDi (Moran et al., 2022), Tilde (Rozis and Skadiņš, 2017), W2C (Majliš, 2011), WAT (Nakazawa et al., 2022), WikiMatrix (Schwenk et al., 2021), Wikipedia (Wikipedia, 2024), Workshop on NER for South and South East Asian Languages (Singh, 2008a), and XLSum (Hasan et al., 2021).
- MADLAD-400 (Kudugunta et al., 2023):
  CommonCrawl (Common Crawl, 2022).

We start with the corpus from Chang et al. (2024a). We then merge the dataset per language with Glot500 for languages that have not yet reached our 1GB maximum (after byte premium scaling). Then, we merge the dataset with MADLAD-400 for languages that have still not reached our 1GB maximum. We also add MADLAD-400 for languages with short average line lengths (less than 25.0 tokens), to make use of MADLAD-400's longer contiguous sequences. To allow comparisons on popular low-resource language evaluations, we exclude FLORES-200 (Costa-jussà et al., 2022) and AmericasNLI (Ebrahimi et al., 2022) from all dataset merging. For each dataset, we exclude languages that contain only Bible data. Because there is likely significant overlap between different dataset sources, we deduplicate repeated sequences of 100 UTF-8 bytes for each language (Lee et al., 2022).

Language codes. To enable dataset merging per language, several datasets must be converted to ISO 639-3 language codes and ISO 15924 script codes. In some cases, this introduces ambiguity because datasets can be labeled as individual language codes (e.g. `quy_latn` for Ayacucho Quechua and `quz_latn` for Cusco Quechua) or as macrolanguage codes (e.g. `que_latn` for Quechua). In these cases, we compile both a macrolanguage dataset and individual language datasets. Datasets labeled with individual codes contribute both to their individual dataset and their umbrella macrolanguage dataset; datasets labeled with macrolanguage codes contribute only to the macrolanguage dataset. For example, we have individual `quy_latn` and `quz_latn` datasets, both of which contribute to a larger `que_latn` dataset, which also contains datasets labeled only with `que_latn`. These ambiguities primarily appear for lower-resource languages.

We also drop several redundant language codes:

- We drop `ory_orya` (Odia) in favor of the macrocode `ori_orya` because `ory_orya` is the only individual language within `ori_orya` for which we have any data.
- For the same reason, we drop `npi_deva` (Nepali) in favor of the macrocode `nep_deva`.
- For the same reason, we drop `swh_latn` (Swahili) in favor of the macrocode `swa_latn`.
- We drop `cmn_hans` (Mandarin) in favor of the macrocode `zho_hans` (Chinese) because the `zho_hans` data is almost entirely in Mandarin. While less specific, `zho_hans` is commonly used by other datasets. For other Chinese languages, see their individual codes (e.g. `yue_hant` for Cantonese). We note that the similar code `zho_hant` (traditional characters) is not primarily Mandarin.
- We drop `hbs_cyrl` and `hbs_latn` (Serbo-Croatian) because we have the individual languages Serbian (`srp_cyrl` and `srp_latn`), Croatian (`hrv_latn`), and Bosnian (`bos_cyrl` and `bos_latn`).
- We drop the deprecated code `ajp_arab` (Levantine Arabic) in favor of `apc_arab`.
- We drop `ber_latn` (Berber) because it is a collective code for distinct (and often not mutually intelligible) languages. We keep the constituent individual languages.
- We drop `nah_latn` (Nahuatl) because it is a collective code for distinct languages. We keep the constituent individual languages.

After merging, we have a dataset of 547GB of text covering 523 language-script combinations (486 unique language codes, 32 unique script codes).

Byte premiums. As described in §3.1, we then scale our dataset sizes by estimated byte premiums (Arnett et al., 2024). A byte premium b for a language L indicates that content-matched (i.e. parallel) text in L takes $b\times$ as many UTF-8 bytes to encode as English. We use the Byte Premium Tool (Arnett et al., 2024) to compute or estimate the byte premium for all of our languages. Byte premiums are pre-computed in the tool for high-resource languages. For each novel low-resource language L, we use the tool (which uses a linear regression) to predict the byte premium for L based on the character entropy for text in L and the script type for L (alphabet, abjad, abugida, or logography), as recommended for low-resource languages in Arnett et al. (2024). Then, we have an estimated byte premium for every language in our dataset. We clip each byte premium to a minimum of 0.70 and a maximum of 5.00; clipping occurs for only three languages (`lzh_hant`, `wuu_hani` $\rightarrow$ 0.70, `mya_mymr` $\rightarrow$ 5.00). As described in §3.1, all of our training datasets (both for tokenizers and for the models themselves) are sampled based on size in bytes after byte premium scaling. We drop languages with less than 5MB of text after byte premium scaling.

Dataset statistics. The resulting 350 Goldfish languages cover five continents, 28 top-level language families (Hammarström et al., 2023), and 32 scripts (writing systems). All languages for which Goldfish models are available are listed in Table 5. We include the language name, ISO 639-3 language code, ISO 15924 script code, estimated byte premium, dataset size after byte premium scaling, dataset size in tokens, and proportion of the dataset from each of our four largest sources. Raw dataset sizes before byte premium scaling can be obtained by multiplying the dataset size after byte premium scaling by the estimated byte premium. Source dataset proportions are reported before deduplication. The reported dataset sizes reflect the dataset for the Goldfish model trained on the maximum amount of data for that language (the **1GB**-dataset Goldfish when available, otherwise the **full**-dataset Goldfish). Reported token counts use the tokenizer for the largest Goldfish model for that language. Datasets can be downloaded at: `https://huggingface.co/datasets/goldfish-models/fish-food`.

A.2. Contamination Analysis

While we aim to exclude the FLORES dataset from our training datasets, there is still the possibility that FLORES could be contaminated inadvertently due to the presence of web data. To check for any potential contamination of FLORES in the datasets,

Hyperparameter	5MB, 10MB	full, 100MB+
Total parameters	39M	125M
Layers	4	12
Embedding size	512	768
Hidden size	512	768
Intermediate hidden	2048	3072
Attention heads	8	12
Attention head size	64	64
Learning rate	1e-4	
Batch size	5MB: 4, 10MB: 8, 100MB: 32, 1GB: 64	
Epochs	10	
Activation function	GELU	
Max sequence length	512	
Position embedding	Absolute	
Learning rate decay	Linear	
Warmup steps	10% of pretraining	
Adam ϵ	1e-6	
Adam β_1	0.9	
Adam β_2	0.999	
Dropout	0.1	
Attention dropout	0.1	

Table 4: Pretraining hyperparameters for Goldfish trained on different dataset sizes (Devlin et al., 2019; Turc et al., 2019; Radford et al., 2018).

we tokenize each FLORES sequence and the entire training dataset for each language that is in both Goldfish and FLORES (204 languages). We compute the total number of times that the first 10 tokens of any FLORES example appears in the training dataset for the language. For 72% of languages, no FLORES examples appear in the training dataset at all. For 98% of languages, less than 10 FLORES examples appear in the training dataset (out of over 2000 FLORES examples total). The only two languages with notable FLORES contamination are `smo_latn` (Samoan; 7155 occurrences of FLORES examples in the training dataset) and `knc_arab` (Central Kanuri; 371 occurrences of FLORES examples in the training dataset).

A.3. Pretraining Details

Architectures. All of our models use the GPT-2 architecture (Radford et al., 2019), changing only the number of layers, attention heads, and embedding sizes as in Turc et al. (2019). For the 100MB-, 1GB-, and full-dataset models, we use the 125M-parameter architecture equivalent to GPT-1 (Radford et al., 2018) (similar to BERT-base and RoBERTa; Devlin et al., 2019; Liu et al., 2019). Because smaller models perform similarly to larger models in low-resource scenarios (Chang et al., 2024a), we use the small model size (39M parameters) from Turc et al. (2019) for the 10MB and 5MB dataset sizes.

Training hyperparameters. Language models are pretrained using the Hugging Face Transformers library (Wolf et al., 2020) and code from Chang and Bergen (2022). We refrain from extensive hyperparameter tuning to avoid biasing our hyperparameters towards English (or any other selected tuning language). Instead, we adopt hyperparameters from previous work with minimal modifications. To match the setup of our models and to prevent overfitting, we select hyperparameters based on models with fairly small training datasets relative to modern standards. Specifically, following BERT (Devlin et al., 2019), we use learning rate 1e-4 for the 125M-parameter models (the same as RoBERTa for small batch sizes; Liu et al., 2019; GPT-1 uses learning rate 2.5e-4; Radford et al., 2018). Based on initial results using randomly-sampled languages, we find that learning rate 1e-4 also works well for the 39M-parameter models; this is in line with Chang et al. (2024a), who find that learning rate 2e-4 works well for small models, and smaller learning rates reduce the speed of any potential overfitting.

Following GPT-1 (most similar to our models; Radford et al., 2018), we use batch size 64 (64×512 = 32K tokens) for the 1GB-dataset models. We find that these larger batch sizes lead to overfitting for small datasets, so we use batch sizes 4, 8, and 32 for 5MB-, 10MB-, and 100MB-dataset models respectively (determined based on initial experiments with randomly-sampled languages). These correspond to batches of 2K, 4K, or 16K tokens. For full-dataset models, we use the batch size that would be used if rounding the dataset size down to 5MB, 10MB, or 100MB (recall that we do not train a full-dataset model when the 1GB dataset is available for a language).

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. In total, Goldfish pretraining takes the equivalent of approximately 15600 A6000 GPU hours. Inference for FLORES perplexities and reasoning benchmarks takes approximately 250 A6000 GPU hours (primarily due to the large multilingual models used for comparison). Dataset merging, deduplication, and tokenization takes approximately 1600 CPU core hours.

A.4. FLORES Evaluation Details

In §4.1, we evaluate the Goldfish models, XGLM 4.5B, XGLM 7.5B, BLOOM 7.1B, MaLA-500 10B, and bigram models on FLORES log-perplexity (negative log-likelihood). To avoid confounds from different tokenizers across models, we compute log-perplexities at the sequence level. For fair comparison with multilingual models that need to determine the input language during the early parts of a sequence, we compute log-perplexity of the second

half s_1 of each sequence given the first half s_0. We then compute the mean over sequences:

$$\text{LogPPL}_{\mathcal{M}} = \text{mean}_s\Big(-\log(P_{\mathcal{M}}(s_1|s_0))\Big) \quad (2)$$

A lower log-perplexity indicates better performance, where $\mathcal{M}$ assigns higher probabilities to ground truth text (FLORES sequences). While imperfect, perplexity does not require annotated text data, it is predictive of performance on a variety of downstream tasks (Xia et al., 2023), and it has been used to measure language model quality in previous work (Kaplan et al., 2020; Hoffmann et al., 2022; Lin et al., 2024).

In detail, the first and second half of each sequence are determined based on number of characters, so the halfway split is the same for all models considered. We round to the nearest token when the halfway split is in the middle of a subword token. Each model $\mathcal{M}$ then assigns some probability $P_{\mathcal{M}}(s_1|s_0)$ regardless of tokenization, except for rounding the halfway point to the nearest token. The probability for any [UNK] (unknown) token is set to random chance $1/v$ where v is the tokenizer vocabulary size.[5] As our final log-perplexity score, we compute the mean negative-log-probability over all FLORES sequences in the target language. Because perplexities generally use geometric means, we use arithmetic means for log-perplexities. The final equation is presented in Equation 2.

The mean FLORES perplexities for each model are reported in Table 1. For Goldfish models, we report the perplexity for the model trained on the largest dataset for the language (i.e. the 1GB-dataset model when available, otherwise the full-dataset model). Perplexities per language for the 5MB-, 10MB-, 100MB-, and 1GB-dataset models specifically are available at https://github.com/tylerachang/goldfish.

Ambiguous or missing languages. Several of the FLORES and Belebele languages are either missing from Goldfish or have multiple possible Goldfish available (e.g. either the macrolanguage que_latn or individual language quy_latn for FLORES language quy_latn). We make the following substitutions:

- taq_tfng → None,
  tzm_tfng → None.
  None of the language models evaluated are trained on these languages, and no Goldfish are trained with the Tifinagh (tfng) script.
- awa_deva → hin_deva,
  kam_latn → kik_latn.
  kas_arab → urd_arab,
  mni_beng → ben_beng,
  nus_latn → din_latn,
  taq_latn → kab_latn.
  Here, we use the closest relative in Goldfish that uses the same script.
- ace_arab → urd_arab,
  arb_latn → mlt_latn,
  ben_latn → hin_latn,
  bjn_arab → urd_arab,
  min_arab → urd_arab,
  npi_latn → hin_latn,
  sin_latn → hin_latn,
  urd_latn → hin_latn.
  These are languages that are missing from Goldfish and that are written in a nonstandard script for the language (e.g. Arabic in Latin script). We use the closest relative in Goldfish that uses that script.
- acm_arab → arb_arab,
  acq_arab → arb_arab,
  aeb_arab → arb_arab,
  ajp_arab → arb_arab,
  als_latn → sqi_latn,
  ars_arab → arb_arab,
  ary_arab → arb_arab,
  ayr_latn → aym_latn,
  azb_arab → aze_arab,
  azj_latn → aze_latn,
  dik_latn → din_latn,
  gaz_latn → orm_latn,
  khk_cyrl → mon_cyrl,
  kmr_latn → kur_latn,
  lvs_latn → lav_latn,
  npi_deva → nep_deva,
  ory_orya → ori_orya,
  pbt_arab → pus_arab,
  plt_latn → mlg_latn,
  quy_latn → que_latn,
  swh_latn → swa_latn,
  uzn_latn → uzb_latn,
  ydd_hebr → yid_hebr,
  yue_hant → zho_hant,
  zsm_latn → msa_latn.
  These languages map to multiple different Goldfish languages or are individual languages within a macrolanguage code included in Goldfish. When the option is available, we use the Goldfish language with more data.

[5] Otherwise, for unseen writing systems (e.g. Tibetan script tibt in XGLM), the probability $P([\text{UNK}]|[\text{UNK}]\ [\text{UNK}]\ ...)$ is very high, resulting in artificially low perplexities. Setting the [UNK] token probabilities to random chance has very little effect on log-perplexity scores except for the scenario of an unseen writing system.

Table 5: Goldfish languages with corresponding dataset sizes. Horizontal lines separate languages with at least 1000MB, 100MB, 10MB, and 5MB of available data.

Language	Language (ISO 639-3)	Script (ISO 15924)	Byte Premium	Scaled MB	Tokens	Dataset Proportions OSCAR NLLB MADLAD-400 Glot500 Other
Afrikaans	afr	latn	1.04	1000.00	239682048	
Amharic	amh	ethi	1.72	1000.00	211767808	
Standard Arabic	arb	arab	1.47	1000.00	196197376	
Azerbaijani	aze	latn	1.30	1000.00	233091584	
Belarusian	bel	cyrl	2.01	1000.00	254138368	
Bengali	ben	beng	2.43	1000.00	194737152	
Bosnian	bos	cyrl	1.15	1000.00	232501760	
Bosnian	bos	latn	0.97	1000.00	228266496	
Bulgarian	bul	cyrl	1.81	1000.00	224346112	
Catalan	cat	latn	1.09	1000.00	238915072	
Czech	ces	latn	1.04	1000.00	206113280	
Welsh	cym	latn	1.03	1000.00	236230144	
Danish	dan	latn	1.02	1000.00	208085504	
German	deu	latn	1.05	1000.00	210817024	
Modern Greek	ell	grek	1.97	1000.00	238704128	
English	eng	latn	1.00	1000.00	213977088	
Esperanto	epo	latn	1.00	1000.00	231384576	
Estonian	est	latn	0.97	1000.00	189518336	
Basque	eus	latn	1.06	1000.00	209921536	
Persian	fas	arab	1.59	1000.00	244359680	
Filipino	fil	latn	1.33	1000.00	274955776	
Finnish	fin	latn	1.06	1000.00	186050560	
French	fra	latn	1.17	1000.00	251415552	
Galician	glg	latn	1.06	1000.00	222080000	
Gujarati	guj	gujr	2.16	1000.00	193794560	
Hausa	hau	latn	1.18	1000.00	277416448	
Hebrew	heb	hebr	1.36	1000.00	192904704	
Hindi	hin	deva	2.37	1000.00	228020736	
Croatian	hrv	latn	0.99	1000.00	219422208	
Hungarian	hun	latn	1.02	1000.00	191089664	
Armenian	hye	armn	1.72	1000.00	203630592	
Indonesian	ind	latn	1.18	1000.00	210432000	
Icelandic	isl	latn	1.15	1000.00	236872704	
Italian	ita	latn	1.07	1000.00	216099840	
Japanese	jpn	jpan	1.32	1000.00	219063296	
Kara-Kalpak	kaa	cyrl	1.92	1000.00	212100608	
Kannada	kan	knda	2.64	1000.00	212683264	
Georgian	kat	geor	4.34	1000.00	354762752	
Kazakh	kaz	cyrl	1.76	1000.00	199970304	
Kirghiz	kir	cyrl	1.96	1000.00	223066112	
Korean	kor	hang	1.29	1000.00	227021824	
Latin	lat	latn	0.88	1000.00	188774912	
Latvian	lav	latn	1.29	1000.00	243401728	
Lithuanian	lit	latn	1.03	1000.00	201228800	
Malayalam	mal	mlym	2.88	1000.00	244708864	
Marathi	mar	deva	2.48	1000.00	206630400	
Macedonian	mkd	cyrl	1.83	1000.00	221346304	
Maltese	mlt	latn	1.09	1000.00	283158528	
Mongolian	mon	cyrl	1.78	1000.00	205737472	
Malay	msa	latn	1.29	1000.00	236371456	
Nepali	nep	deva	2.63	1000.00	215368192	

Dutch	nld	latn	1.05	1000.00	216978432	
Norwegian Bokmål	nob	latn	1.00	1000.00	205949952	
Norwegian	nor	latn	1.13	1000.00	255482880	
Panjabi	pan	guru	2.22	1000.00	215775232	
Iranian Persian	pes	arab	1.60	1000.00	215946240	
Polish	pol	latn	1.08	1000.00	216235008	
Portuguese	por	latn	1.10	1000.00	225242112	
Pushto	pus	arab	1.59	1000.00	237871616	
Romanian	ron	latn	1.12	1000.00	230 580224	
Russian	rus	cyrl	1.82	1000.00	220467712	
Sinhala	sin	sinh	2.45	1000.00	233098752	
Slovak	slk	latn	1.04	1000.00	211206144	
Slovenian	slv	latn	0.97	1000.00	198052864	
Somali	som	latn	1.42	1000.00	302652928	
Spanish	spa	latn	1.08	1000.00	221790720	
Albanian	sqi	latn	1.34	1000.00	274664448	
Serbian	srp	cyrl	1.42	1000.00	184423424	
Serbian	srp	latn	0.83	1000.00	207482368	
Swahili	swa	latn	1.26	1000.00	260033024	
Swedish	swe	latn	1.02	1000.00	206359552	
Tamil	tam	taml	2.73	1000.00	200523264	
Tatar	tat	cyrl	1.85	1000.00	232933888	
Telugu	tel	telu	2.62	1000.00	209365504	
Tajik	tgk	cyrl	1.75	1000.00	216990208	
Tagalog	tgl	latn	1.12	1000.00	245370880	
Thai	tha	thai	2.74	1000.00	205872640	
Turkish	tur	latn	1.04	1000.00	186848768	
Ukrainian	ukr	cyrl	1.75	1000.00	215392768	
Urdu	urd	arab	1.71	1000.00	247899648	
Uzbek	uzb	latn	1.23	1000.00	261058560	
Vietnamese	vie	latn	1.35	1000.00	262306304	
Chinese	zho	hans	0.94	1000.00	206204416	
Irish	gle	latn	1.98	976.70	404823040	
Kurdish	kur	arab	1.57	902.39	196483584	
Standard Malay	zsm	latn	1.14	859.52	185929728	
Central Kurdish	ckb	arab	1.65	838.87	190565888	
Kinyarwanda	kin	latn	1.13	810.96	193561088	
Haitian	hat	latn	0.97	775.80	185333248	
Odia	ori	orya	2.60	774.55	165528576	
Zulu	zul	latn	1.16	764.14	199965696	
Burmese	mya	mymr	5.00	762.14	315374592	
Central Khmer	khm	khmr	3.90	742.37	235559424	
Malagasy	mlg	latn	1.27	720.80	210497024	
Kurdish	kur	latn	1.29	685.53	189872128	
Dhivehi	div	thaa	2.00	634.02	114510336	
Shona	sna	latn	1.12	608.11	151712256	
Luxembourgish	ltz	latn	1.23	579.07	160200192	
Sundanese	sun	latn	1.10	577.96	142266368	
Scottish Gaelic	gla	latn	0.99	558.84	123736064	
Cebuano	ceb	latn	1.11	540.21	140301312	
Lao	lao	laoo	2.71	532.98	124077056	
Uzbek	uzb	cyrl	1.98	525.51	110868992	
Yoruba	yor	latn	1.37	502.55	155829248	
Norwegian Nynorsk	nno	latn	1.03	498.93	116016128	
Xhosa	xho	latn	1.20	477.36	127885824	
Western Frisian	fry	latn	1.23	472.81	133072384	
Javanese	jav	latn	1.15	465.58	115332096	
Sindhi	snd	arab	1.59	459.14	114626048	
Maori	mri	latn	1.18	450.17	136011776	
Yiddish	yid	hebr	1.55	446.04	85695488	
Nyanja	nya	latn	1.21	444.13	112440832	

Corsican	cos	latn	1.18	414.00	126150656	
Faroese	fao	latn	1.16	400.34	96587776	
Bashkir	bak	cyrl	2.27	398.36	118369280	
Uighur	uig	arab	2.31	397.21	104039936	
Igbo	ibo	latn	1.35	388.31	119706112	
Modern Greek	ell	latn	1.24	376.42	92225536	
Occitan	oci	latn	1.01	375.38	99783680	
Plateau Malagasy	plt	latn	1.15	370.58	97517568	
Assamese	asm	beng	2.53	348.88	77216256	
Hmong	hmn	latn	1.19	345.97	100051968	
Tosk Albanian	als	latn	1.17	336.30	87609344	
Southern Sotho	sot	latn	1.17	332.91	94144000	
Samoan	smo	latn	1.18	314.93	101910016	
Azerbaijani	aze	arab	1.20	267.26	56526848	
Hawaiian	haw	latn	1.11	260.95	86747136	
Chuvash	chv	cyrl	1.80	256.36	84293120	
Papiamento	pap	latn	1.00	255.51	60037632	
Tigrinya	tir	ethi	1.76	252.98	56515072	
Asturian	ast	latn	1.75	225.68	93333504	
Southern Pashto	pbt	arab	1.74	225.11	60608000	
Central Kanuri	knc	arab	2.50	221.65	237422592	
Lushai	lus	latn	1.17	213.03	62735360	
Northern Uzbek	uzn	cyrl	2.01	208.92	44960768	
Yakut	sah	cyrl	1.88	206.06	47289344	
Ancient Greek	grc	grek	1.77	205.45	47620608	
Turkmen	tuk	latn	1.79	186.44	57201664	
Chinese	zho	hant	0.99	177.32	42692096	
Waray	war	latn	1.09	175.25	48998912	
Kara-Kalpak	kaa	latn	1.23	165.22	38767104	
Breton	bre	latn	1.01	163.11	43437056	
Dari	prs	arab	1.66	162.70	37549568	
Venetian	vec	latn	1.00	150.70	40523776	
North Azerbaijani	azj	latn	1.08	149.82	27041792	
Northern Uzbek	uzn	latn	1.65	145.59	52049408	
Limburgan	lim	latn	1.00	142.31	39700480	
Kalaallisut	kal	latn	1.34	140.44	30082048	
Quechua	que	latn	1.21	139.38	40595968	
Oromo	orm	latn	1.26	137.90	39742976	
Ganda	lug	latn	1.22	132.42	37459968	
Tibetan	bod	tibt	2.62	131.94	23463424	
Hindi	hin	latn	1.26	131.86	37683712	
Swiss German	gsw	latn	1.14	128.81	38605824	
Ayacucho Quechua	quy	latn	1.16	123.58	34850816	
Lombard	lmo	latn	0.94	123.24	35603456	
Egyptian Arabic	arz	arab	1.55	122.38	30322176	
Western Panjabi	pnb	arab	1.41	121.58	30110208	
Eastern Yiddish	ydd	hebr	1.81	120.20	28306432	
Sanskrit	san	deva	2.54	119.34	31856128	
Sicilian	scn	latn	1.04	113.80	32010752	
Halh Mongolian	khk	cyrl	1.80	108.25	23605760	
South Azerbaijani	azb	arab	1.49	107.56	26922496	
Walloon	wln	latn	1.22	102.32	29091328	
Tswana	tsn	latn	1.17	101.85	31488512	
Gujarati	guj	latn	1.19	101.60	24635392	
Gilaki	glk	arab	1.68	98.73	25519104	
Iloko	ilo	latn	1.08	97.44	25450496	
Tetum	tet	latn	1.40	96.03	28032512	
Banjar	bjn	latn	1.17	93.17	25012224	
Rundi	run	latn	1.12	90.59	23721984	
Romansh	roh	latn	1.27	86.73	23623680	
Chechen	che	cyrl	1.83	86.11	23590400	

West Central Oromo	gaz	latn	1.33	79.04	25565184	
Yue Chinese	yue	hant	0.86	78.42	16084992	
Low German	nds	latn	1.14	75.35	20312064	
Minangkabau	min	latn	0.95	75.07	17732608	
Inuktitut	iku	cans	2.16	74.41	13798400	
Tsonga	tso	latn	1.21	71.85	21684224	
Achinese	ace	latn	1.24	71.09	21666816	
Tuvinian	tyv	cyrl	1.86	68.39	15576576	
Northern Sami	sme	latn	1.27	66.64	15802880	
Ewe	ewe	latn	1.08	63.27	18470400	
Twi	twi	latn	1.03	62.79	18900480	
Standard Estonian	ekk	latn	0.99	61.41	12375552	
Guarani	grn	latn	0.99	60.38	15366656	
Pedi	nso	latn	1.12	59.40	17516544	
Northern Kurdish	kmr	latn	1.03	53.71	12299264	
Udmurt	udm	cyrl	1.74	51.77	10932736	
Akan	aka	latn	1.57	49.51	22551040	
Mari (Russia)	chm	cyrl	1.76	49.43	11290624	
Mongolian	mon	latn	1.18	49.21	12692480	
Lingala	lin	latn	1.14	47.33	13213184	
Crimean Tatar	crh	latn	1.31	47.20	12994560	
Zaza	zza	latn	1.20	46.78	14813184	
Kabyle	kab	latn	1.03	45.19	14035456	
Min Nan Chinese	nan	latn	1.15	44.38	16624128	
Scots	sco	latn	1.19	42.97	12578304	
Aragonese	arg	latn	1.19	42.82	12469760	
Maithili	mai	deva	2.39	41.73	11159040	
Fon	fon	latn	1.54	40.84	13993984	
Buriat	bua	cyrl	1.70	39.10	8951808	
Ossetian	oss	cyrl	1.85	38.60	14059008	
Pampanga	pam	latn	1.19	38.14	11270656	
Dimli	diq	latn	0.96	37.98	9935872	
Wolof	wol	latn	1.08	37.32	12005888	
Tedim Chin	ctd	latn	1.30	37.10	11405824	
Tumbuka	tum	latn	1.21	36.69	9842688	
Pangasinan	pag	latn	1.04	36.43	10441728	
Fijian	fij	latn	1.21	35.48	10642944	
Standard Latvian	lvs	latn	1.21	35.42	8333312	
Bemba	bem	latn	1.16	35.35	10177024	
Kabardian	kbd	cyrl	1.78	34.89	9802752	
Luo	luo	latn	1.04	34.50	9859072	
Hakha Chin	cnh	latn	1.32	33.20	10364928	
Hiligaynon	hil	latn	1.35	32.12	9034752	
Balinese	ban	latn	1.27	31.84	9161216	
Aymara	aym	latn	1.21	30.74	9201152	
Avaric	ava	cyrl	1.94	30.73	8009728	
Central Aymara	ayr	latn	1.10	28.37	7641088	
Fiji Hindi	hif	latn	1.28	28.00	8768000	
Ligurian	lij	latn	1.14	27.89	8498176	
Eastern Mari	mhr	cyrl	1.81	27.86	6580224	
Bavarian	bar	latn	1.13	27.68	7961600	
Silesian	szl	latn	1.07	27.04	7593472	
Russian	rus	latn	1.18	26.62	7373824	
Ido	ido	latn	1.18	26.18	7369216	
Russia Buriat	bxr	cyrl	1.59	25.38	6060544	
Abkhazian	abk	cyrl	2.01	25.24	6408192	
Sardinian	srd	latn	1.11	24.71	6834176	
Nigerian Pidgin	pcm	latn	0.95	24.62	5281280	
Wu Chinese	wuu	hani	0.70	24.53	4112384	
Fulah	ful	latn	1.26	24.03	7806464	
Bhojpuri	bho	deva	2.52	23.74	6156800	

Betawi	bew	cyrl	1.74	23.52	5288960	
Volapük	vol	latn	1.13	21.39	6030336	
Nigerian Fulfulde	fuv	latn	1.11	21.23	6159872	
Karachay-Balkar	krc	cyrl	1.87	21.02	4627456	
Swati	ssw	latn	1.14	20.97	5566976	
Luba-Lulua	lua	latn	1.19	20.82	6322688	
Friulian	fur	latn	1.07	20.72	5487616	
Khasi	kha	latn	1.30	20.56	6209536	
Telugu	tel	latn	1.28	20.02	5266432	
Iban	iba	latn	1.30	19.98	5278208	
Bikol	bik	latn	1.27	19.26	5440512	
Interlingua	ina	latn	1.24	19.15	5581824	
Latgalian	ltg	latn	1.00	18.70	4046848	
Komi	kom	cyrl	1.61	18.20	4716032	
Querétaro Otomi	otq	latn	1.25	17.48	5702656	
Tonga (Tonga Islands)	ton	latn	1.27	17.46	6237184	
Azerbaijani	aze	cyrl	1.82	17.12	3627008	
Dargwa	dar	cyrl	2.02	16.99	4506624	
Erzya	myv	cyrl	1.77	16.81	3851776	
Piemontese	pms	latn	1.23	16.75	5307904	
Tok Pisin	tpi	latn	1.18	16.61	5102592	
Umbundu	umb	latn	1.17	16.12	4743168	
Sango	sag	latn	1.16	15.87	4929024	
Kabuverdianu	kea	latn	0.78	15.74	3247616	
Adyghe	ady	cyrl	1.81	15.22	4124160	
Literary Chinese	lzh	hant	0.70	15.20	2767872	
Gulf Arabic	afb	arab	1.37	14.25	3247616	
Falam Chin	cfm	latn	1.32	14.09	4315648	
Kabiyè	kbp	latn	1.44	13.93	4698624	
Bambara	bam	latn	1.26	12.84	4511744	
Kachin	kac	latn	1.35	12.74	4453888	
Newari	new	deva	2.56	12.44	2927616	
Syriac	syr	syrc	1.41	12.17	2641408	
Chokwe	cjk	latn	1.17	12.10	3622400	
Dyula	dyu	latn	1.15	11.94	3849216	
Betawi	bew	latn	1.30	11.84	3186176	
Venda	ven	latn	1.30	11.82	3268608	
Dinka	din	latn	1.24	11.69	4125696	
Shan	shn	mymr	2.82	11.66	2238976	
Southern Altai	alt	cyrl	1.86	11.65	2694144	
Southwestern Dinka	dik	latn	1.12	11.61	3753984	
Goan Konkani	gom	deva	1.74	11.50	2219520	
Sranan Tongo	srn	latn	1.06	11.47	3098112	
Yucateco	yua	latn	1.24	11.41	3645440	
Kongo	kon	latn	1.23	11.32	3549184	
Kimbundu	kmb	latn	1.13	11.09	3359744	
Kumyk	kum	cyrl	1.96	11.04	2208768	
Buginese	bug	latn	1.23	10.72	3269632	
Goan Konkani	gom	latn	1.21	10.38	2806784	
Mossi	mos	latn	1.14	10.37	3537920	
Upper Sorbian	hsb	latn	1.12	10.31	2503680	
Lak	lbe	cyrl	2.01	10.24	2470912	
North Ndebele	nde	latn	0.97	10.17	1766912	
Central Kanuri	knc	latn	1.18	10.07	3433472	
Ingush	inh	cyrl	1.70	9.59	2764800	
Zapotec	zap	latn	1.08	9.58	2395136	
Central Bikol	bcl	latn	1.22	9.49	2638336	
Lezghian	lez	cyrl	1.83	9.38	2358784	
Kituba	mkw	cyrl	1.81	9.37	2266112	
Cusco Quechua	quz	latn	1.30	9.32	2070528	
Bishnupriya	bpy	beng	2.33	9.29	2019328	

Mam	mam	latn	1.34	9.27	3580416	
Magahi	mag	deva	2.56	9.08	2488832	
Tzotzil	tzo	latn	1.49	9.02	3463680	
Tamil	tam	latn	1.27	9.00	2260992	
Western Mari	mrj	cyrl	1.51	8.74	1812992	
Brunei Bisaya	bsb	latn	1.31	8.69	2460672	
Chhattisgarhi	hne	deva	2.17	8.61	2106880	
Luba-Katanga	lub	latn	1.30	8.61	2269184	
Kaqchikel	cak	latn	1.82	8.51	4157952	
Santali	sat	olck	2.80	8.49	2224128	
Vlaams	vls	latn	1.21	8.49	2484736	
Kikuyu	kik	latn	1.29	8.36	2418176	
Mirandese	mwl	latn	1.24	8.12	2293760	
Isoko	iso	latn	1.48	8.11	2638336	
Uighur	uig	latn	1.19	7.88	1662976	
Dzongkha	dzo	tibt	3.26	7.70	2019328	
Bashkir	bak	latn	1.19	7.53	1793024	
Dombe	dov	latn	0.99	7.43	1389056	
Madurese	mad	latn	1.29	7.29	2044416	
Levantine Arabic	apc	arab	1.47	7.06	1687040	
Pohnpeian	pon	latn	0.90	7.02	1412608	
Kashmiri	kas	deva	2.53	6.96	1990656	
Paite Chin	pck	latn	1.32	6.94	2163712	
Veps	vep	latn	1.17	6.89	1751552	
Boko (Benin)	bqc	latn	0.98	6.80	1806336	
Neapolitan	nap	latn	1.23	6.73	2123776	
Manx	glv	latn	1.22	6.63	1939968	
Nande	nnb	latn	1.31	6.49	1764352	
Batak Toba	bbc	latn	1.33	6.48	1846784	
Malayalam	mal	latn	1.27	6.38	1556480	
Tiv	tiv	latn	1.31	6.32	2119168	
Cornish	cor	latn	1.22	6.31	1936896	
Khakas	kjh	cyrl	1.93	6.17	1271808	
Moksha	mdf	cyrl	1.71	6.17	1302016	
Kalmyk	xal	cyrl	1.72	6.05	1474048	
Guerrero Nahuatl	ngu	latn	1.44	5.99	1508864	
Klingon	tlh	latn	1.14	5.91	1741312	
Crimean Tatar	crh	cyrl	1.89	5.86	1265664	
Makhuwa-Meetto	mgh	latn	1.11	5.77	1251328	
Sanskrit	san	latn	0.97	5.72	1164800	
Northern Frisian	frr	latn	1.17	5.68	1594368	
Eastern Balochi	bgp	latn	1.29	5.64	1735680	
Carpathian Romani	rmc	latn	1.02	5.61	1241600	
Georgian	kat	latn	1.20	5.57	1422336	
Old English	ang	latn	1.29	5.47	1671168	
Kedah Malay	meo	latn	1.28	5.44	1670656	
Mingrelian	xmf	geor	2.51	5.44	1367040	
Tulu	tcy	knda	2.67	5.29	1210368	
Tandroy-Mahafaly Malagasy	tdx	latn	1.00	5.23	1303552	
Komi-Zyrian	kpv	cyrl	1.67	5.19	1355776	
Lingua Franca Nova	lfn	latn	1.30	5.12	1593344	
Ditammari	tbz	latn	1.33	5.12	1868800	
Nzima	nzi	latn	1.42	5.07	1514496	
Rusyn	rue	cyrl	1.56	5.03	1160704	
Eastern Huasteca Nahuatl	nhe	latn	1.49	5.02	1268224	