Goldfish: Monolingual Language Models for 350 Languages

Tyler A. Chang[1,2], **Catherine Arnett**[3], **Zhuowen Tu**[1], **Benjamin K. Bergen**[1]
[1]Department of Cognitive Science
[2]Halıcıoğlu Data Science Institute
[3]Department of Linguistics
University of California San Diego
{tachang, ccarnett, ztu, bkbergen}@ucsd.edu

Abstract

For many low-resource languages, the only available language models are large multilingual models trained on many languages simultaneously. However, using FLORES perplexity as a metric, we find that these models perform worse than bigrams for many languages (e.g. 24% of languages in XGLM 4.5B; 43% in BLOOM 7.1B). To facilitate research that focuses on low-resource languages, we pre-train and release Goldfish, a suite of monolingual autoregressive Transformer language models up to 125M parameters for 350 languages. The Goldfish reach lower FLORES perplexities than BLOOM, XGLM, and MaLA-500 on 98 of 204 FLORES languages, despite each Goldfish model being over 10× smaller. However, the Goldfish significantly underperform larger multilingual models on reasoning benchmarks, suggesting that for low-resource languages, multilinguality primarily improves general reasoning abilities rather than basic text generation. We release models trained on 5MB (350 languages), 10MB (288 languages), 100MB (166 languages), and 1GB (83 languages) of text data where available. The Goldfish models are available as baselines, fine-tuning sources, or augmentations to existing models in low-resource NLP research, and they are further useful for crosslinguistic studies requiring maximally comparable models across languages.

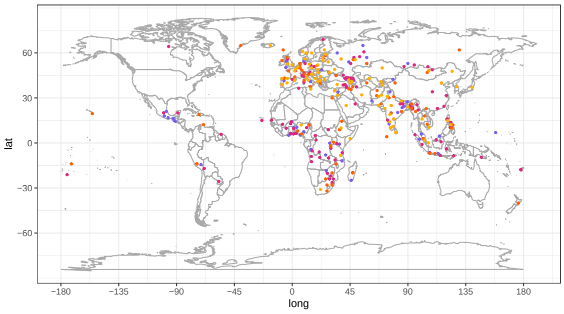
1 Introduction

Language modeling research in low-resource languages often relies on large multilingual models trained on many languages simultaneously (Conneau et al., 2020; 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. 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., 2020).

To address this lack of available models, we introduce Goldfish, a suite of over 1000 monolingual language models for 350 diverse languages.[1] The models 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. 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). However, despite better perplexities, the Goldfish underperform larger multilingual models on reasoning benchmarks, suggesting that multilingual pre-training may benefit abstract reasoning capabilities over more basic grammatical text generation (§5).

Finally, to enable comparisons across languages, we release monolingual models trained on comparable dataset sizes for all languages: 5MB, 10MB, 100MB, and 1GB when available, after accounting for the fact that languages require different numbers of UTF-8 bytes to encode comparable content (Arnett et al., 2024). These Goldfish serve as baselines, allowing results in diverse languages to be situated relative to comparable models. They can also be used as source models for fine-tuning or to enhance larger multilingual models in areas where those models fall short (§4). Models and code are available at `https://huggingface.co/goldfish-models`.

[1]The 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).



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). 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).

2 Related Work

Low resource language modeling often leverages multilingual pre-training, where a model is trained on multiple languages simultaneously (Pires et al., 2019; Conneau et al., 2020). 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., 2023). 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., 2020; Chang et al., 2023).

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 pre-training 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). Unfortunately, comparable monolingual language models across many diverse languages have yet to be studied or released.

3 Models and Datasets

We introduce the Goldfish models, a suite of 1154 monolingual Transformer language models pre-trained 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. (2023), Glot500 (Imani et al., 2023), and MADLAD-400 (Kudugunta et al., 2023) per language. To facilitate fair evaluations, we hold out FLORES-200 and AmericasNLI from all datasets (Costa-jussà et al., 2022; Ebrahimi et al., 2022). We deduplicate repeated sequences of 100 UTF-8 bytes and drop languages with only Bible data. Full dataset details are in §A.1.

To sample pre-training 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× 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 mea-

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 log-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.

suring all datasets in units of "equivalent" English text bytes. We sample datasets to train monolingual language models on **5MB** (350 languages), **10MB** (288 languages), **100MB** (166 languages), and **1GB** (83 languages) 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.

3.2 Architectures and Pre-Training

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., 2023), 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.2.

We tokenize each dataset using using a monolingual SentencePiece tokenizer (Kudo and Richardson, 2018) trained on that dataset size, limiting tokenizer training text to 100MB after byte premium scaling. Following Liu et al. (2019), we use vocabulary size 50K and a maximum sequence length of 512 tokens for all models. We train each language model on 10 epochs of its corresponding dataset.[3] Pre-training details, compute costs, and all available models are reported in §A.2.

4 FLORES Log-Perplexity Evaluations

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 tokenization confounds from computing log-perplexity per token, we compute log-perplexity per FLORES sequence. 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 models 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 simple bigram models trained on the Goldfish datasets.[4] In all cases, we use the Goldfish model trained on the maximum amount of data in each language (maximum 1GB).

FLORES log-perplexity results. The Goldfish reach lower log-perplexities than all four comparison models on 98 of the 204 FLORES languages. Average log-perplexities for the 100 FLORES languages included in both XGLM 4.5B and MaLA-

[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]Multiple epochs of pre-training is beneficial in data-constrained scenarios (Muennighoff et al., 2023), but we find that more than 10 epochs of training leads to overfitting for extremely small datasets (e.g. 5MB).

[4]Bigram and perplexity implementation details in §A.3.

	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.

	# Langs	Chance	Goldfish	XGLM 4.5B	XGLM 7.5B	BLOOM 7.1B	MaLA-500 10B
Belebele	121	25.0	28.2	30.1	**30.6**	30.2	**30.6**
XCOPA	11	50.0	54.9	57.9	**60.6**	56.9	55.6
XStoryCloze	10	50.0	52.5	57.1	**59.9**	58.2	55.7

Table 3: Reasoning benchmark accuracies averaged over non-English languages. Despite better perplexities, the Goldfish perform significantly worse than larger multilingual models on reasoning.

500 are reported in Table 1 (excluding XGLM 7.5B and BLOOM 7.1B there because they are trained on far fewer languages). On average, the Goldfish reach 13% lower log-perplexities than XGLM 4.5B, and 11% lower than MaLA-500 10B.

To ensure that these results are not driven by a small subset of specific languages, in Table 2 we 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. Log-perplexities for individual languages and models are reported in Table 5.

5 Multilingual Reasoning Benchmarks

Because FLORES perplexities are not necessarily reflective of complex capabilities in language models, we also evaluate Goldfish, XGLM, BLOOM, and MaLA-500 (as in §4) on non-English Belebele (121 languages, reading comprehension; Bandarkar et al., 2024), XCOPA (11 languages, commonsense; Ponti et al., 2020), and XStoryCloze (10 languages, story commonsense; Lin et al., 2022). All models are evaluated zero shot with no fine-tuning. Evaluation task details are in §A.4.

Results for all three reasoning tasks are reported in Table 3. Although all models perform quite poorly (close to chance accuracy), the Goldfish perform substantially worse than the multilingual models.[5] This indicates that the combination of larger datasets and model sizes in multilingual pre-training can allow language models to develop reasoning capabilities in specific languages, even when perplexities in those languages remain high. For example, XGLM 7.5B has worse perplexities than Goldfish for 82 Belebele languages (in fact, worse than bigrams for 77 languages), but it outperforms Goldfish on Belebele (reading comprehension) for 56 of those languages. This is in stark contrast with monolingual language models, which generally must reach low perplexities and acquire basic grammatical capabilities before developing reasoning abilities (Liu et al., 2021; Choshen et al., 2022; Xia et al., 2023; Chang et al., 2024). Intuitively, it may be that abstract reasoning patterns are often more language-agnostic than grammatical text generation, and thus multilingual pre-training primarily benefits the former.

6 Conclusion

We pre-train and release Goldfish, a suite of over 1000 monolingual language models for 350 languages. The Goldfish achieve perplexities that are competitive with, and on average lower than, state-of-the-art multilingual language models across languages. However, they underperform large multilingual models on reasoning tasks; in low-resource languages, it appears that multilingual pre-training facilitates nontrivial reasoning capabilities despite extremely poor perplexities. We publicly release all Goldfish models to be used as comparable baselines, fine-tuning sources, or augmentations to larger models (e.g. cross-lingual experts; Blevins et al., 2024) in future low-resource NLP research.

[5] It is unlikely that this effect is due to model size alone; the Goldfish models (125M parameters) have easily enough capacity for their maximum of 1GB of text data.

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., 2020; Chang et al., 2023). 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., 2023). 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) and three reasoning benchmarks (§5). These are some of the only evaluations that can be used for autoregressive language models in many languages, but they have significant limitations. Perplexity is not necessarily predictive of grammatical text generation (Hu et al., 2020) or complex reasoning capabilities (Levy et al., 2024), but it still provides reasonable signal for model performance (Xia et al., 2023) and it is often used to roughly quantify language model quality (Kaplan et al., 2020; Hoffmann et al., 2022). On the other hand, reasoning benchmarks require annotated datasets and thus often cover fewer languages. One notable exception is Belebele (121 non-English languages; Bandarkar et al., 2024), but even large state-of-the-art models perform quite poorly on Belebele without tuning or few-shot prompting (§5). Thus, our evaluations of model reasoning are not entirely 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"). We hope that tractable evaluation datasets with broad language coverage will become increasingly available in the future.

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. We do not redistribute the data itself. Furthermore, 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.

Acknowledgments

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

References

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

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).

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).

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

AI4Bharat. 2023. AI4Bharat. Dataset.

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).

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.

Anuvaad. 2023. Anuvaad project. Dataset.

Catherine Arnett, Tyler A Chang, and Benjamin K Bergen. 2024. A bit of a problem: Measurement disparities in dataset sizes across languages. *arXiv preprint arXiv:2403.00686*.

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.

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

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

Terra Blevins, 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.

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 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).

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.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. 2023. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. *arXiv*.

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. 2024. 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.

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.

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, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. 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.

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

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

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

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).

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

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

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.

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.

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.

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.

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.

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. 2020. 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.

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.

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.

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

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

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

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).

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*.

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.

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

Lingala Songs. 2023. Lingala song lyrics. Dataset.

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.

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.

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.

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.

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*.

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.

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.

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.

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.

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.

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.

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.

Anil Kumar Singh. 2008. 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*.

Stanford. 2023. Stanford nlp group datasets. Dataset.

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.

Daniela Teodorescu, Josie Matalski, Delaney Lothian, Denilson Barbosa, and Carrie Demmans Epp. 2022. Cree corpus: A collection of nêhiyawêwin resources. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6354–6364. Association for Computational Linguistics.

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

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

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

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

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.

Wikipedia. 2024. Wikipedia.

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.

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

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

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.

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.

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

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.

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. (2023), 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. (2023):
  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., 2020), 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, 2008), 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. (2023). 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.

Additionally, we 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 6. 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. All dataset statistics can be downloaded at `https://github.com/tylerachang/goldfish`.

Hyperparameter	**5MB,10MB**	**100MB,1GB,full**
Total parameters	39M	125M
Layers	4	12
Embedding size	512	768
Hidden size	512	768
Intermediate hidden size	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 pre-training	
Adam ϵ	1e-6	
Adam β_1	0.9	
Adam β_2	0.999	
Dropout	0.1	
Attention dropout	0.1	

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

A.2 Pre-Training Details

As described in §3.2, 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). Full hyperparameters are reported in Table 4.

Tokenizers. All tokenizers are trained with vocabulary size 50K (Liu et al., 2019) on the same dataset size as their corresponding model (including byte premium scaling). We use SentencePiece tokenizers (Kudo and Richardson, 2018) with training text randomly sampled from the dataset for the desired language. To avoid memory errors, we limit tokenizer training text to 100MB after byte premium scaling. 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.[6]

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., 2023), 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 pre-trained 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. (2023), who find that learning rate 2e-4 works well for small models, and smaller learning rates reduce the speed of any potential overfitting.

We train each model for 10 epochs of the training data; multiple epochs of pre-training is beneficial in data-constrained scenarios (Muennighoff et al., 2023), but pre-training on more than 10 epochs often leads to overfitting (increases in eval loss) in the 5MB scenarios. For batch sizes, 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).

Compute costs. All language model pre-training 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., 2020; 3.14×10^{23} FLOPs). Models are each trained on one NVIDIA GeForce GTX TITAN X, GeForce RTX 2080 Ti, TITAN Xp, Quadro P6000, RTX A4500, RTX A5000, or RTX A6000 GPU. In total, Goldfish pre-training 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.3 FLORES Evaluation Details

In §4, 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). For each FLORES sequence s, we compute the probability of the second half s_1 of the sequence given the first half s_0. The first and second half 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 vocab-

[6] When de-tokenized, the tokenized datasets result in slightly smaller datasets than the original text datasets, because the tokenizer truncates lines to create 512-token sequences. All reported dataset sizes account for this truncation.

ulary size.[7] 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 1.

FLORES log-perplexities for all models and languages are reported in Table 5. For Goldfish models, we report the log-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). Log-perplexities of the 5MB-, 10MB-, 100MB-, and 1GB-dataset models specifically are available at `https://github.com/tylerachang/goldfish`.

Bigram model details. For each FLORES language, we train a bigram model on the entire Goldfish dataset for that language, up to 1GB after byte premium scaling §3.1. The 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. The tokenizer is the same as the Goldfish tokenizer for that dataset (i.e. the 1GB-dataset model when available, or the full-dataset model). 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.

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.

[7] Otherwise, for unseen writing systems (e.g. Tibetan script `tibt` in XGLM), the probability P([UNK]|[UNK] [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.

A.4 Reasoning Task Details

In §5, we evaluate the Goldfish models, XGLM 4.5B, XGLM 7.5B, BLOOM 7.1B, and MaLA-500 10B on:

- Non-English Belebele (121 languages, reading comprehension; Bandarkar et al., 2024). For languages that are ambiguous or missing from Goldfish, we use the same language code mapping as in §A.3. Each Belebele example consists of a passage, a question, and four candidate answers. We evaluate model accuracy in selecting the correct answer by computing text probabilities for each "[passage] [question] [answer_option]". No model exceeds 41% accuracy for any language (random chance 25%).

- XCOPA (11 languages, commonsense reasoning; Ponti et al., 2020). Each example consists of a premise sentence and two possible causes or effects (i.e. answer options). We use the task format and evaluation implementation in Gao et al. (2023). This selects answers based on a model's computed text probabilities for each "[premise] [connecting_word] [cause/effect_option]", where the connecting word is the translation of "because" (for causes) or "therefore" (for effects).

- Non-English XStoryCloze (10 languages, story commonsense; Lin et al., 2022). Each example consists of a context story and two possible story completions. We use the task format and evaluation implementation in Gao et al. (2023). This selects answers based on a model's computed text probabilities for each "[story_context] [completion_option]".

All models are evaluated zero shot with no fine-tuning. Results per language are available at `https://github.com/tylerachang/goldfish`.

Table 5: FLORES log-perplexity score (↓) for each model and FLORES language. Parentheses indicate that the model is not trained specifically on that language.

Language	FLORES log-perplexity					
	Goldfish	**Bigram**	**XGLM 4.5B**	**XGLM 7.5B**	**BLOOM 7.1B**	**MaLA-500 10B**
ace_arab	(287.55)	(365.59)	(260.50)	(263.32)	**(232.76)**	(251.92)
ace_latn	144.62	169.33	(202.67)	(208.64)	(198.50)	**133.10**
acm_arab	(96.60)	(124.48)	(93.25)	(88.91)	**(85.28)**	102.65
acq_arab	(95.54)	(124.94)	(92.14)	(88.09)	**(82.88)**	(105.38)
aeb_arab	(116.01)	(140.30)	(113.84)	(108.98)	**(102.76)**	(120.88)
afr_latn	79.88	115.57	**79.54**	(162.73)	(153.05)	85.61
ajp_arab	(98.92)	(125.28)	(96.25)	(91.49)	**(85.73)**	103.56
aka_latn	132.48	162.51	(234.93)	(239.68)	187.66	**128.37**
als_latn	77.28	119.91	**76.37**	(220.78)	(178.58)	89.49
amh_ethi	**83.36**	111.87	110.54	(266.96)	(195.14)	108.99
apc_arab	173.18	179.37	(99.97)	(94.47)	**(87.38)**	106.58
arb_arab	82.43	117.61	79.33	75.07	**69.24**	96.84
arb_latn	(245.97)	(346.80)	**211.13**	(226.14)	(221.60)	(229.07)
ars_arab	(83.68)	(118.50)	(80.72)	(76.51)	**(70.75)**	(98.25)
ary_arab	(128.66)	(155.55)	(131.80)	(125.88)	**(114.30)**	123.30
arz_arab	116.98	146.50	(98.72)	(92.96)	**(87.42)**	112.84
asm_beng	**93.78**	118.79	135.60	(227.40)	113.74	108.20
ast_latn	86.86	118.36	(113.30)	(112.31)	(99.39)	**82.54**
awa_deva	(128.70)	(169.05)	(148.81)	(135.84)	**(123.28)**	(141.50)
ayr_latn	**123.33**	148.81	(239.30)	(231.82)	(228.59)	146.85
azb_arab	**154.24**	185.83	163.32	(218.12)	(225.63)	165.85
azj_latn	**74.28**	106.31	75.33	(183.10)	(185.84)	84.13
bak_cyrl	**79.24**	108.34	(259.93)	(272.19)	(209.63)	92.72
bam_latn	158.88	175.25	188.31	(212.45)	203.82	**143.14**
ban_latn	121.16	137.57	(154.09)	(183.02)	(188.39)	**114.41**
bel_cyrl	**79.50**	124.53	81.13	(226.99)	(211.85)	91.22
bem_latn	**150.39**	174.20	(218.51)	(188.14)	(237.98)	158.61
ben_beng	80.71	109.08	90.55	79.78	**76.54**	94.62
bho_deva	121.88	138.45	(168.84)	(149.88)	**(120.97)**	(130.48)
bjn_arab	(297.57)	(383.82)	(260.05)	(261.94)	**(238.42)**	(257.15)
bjn_latn	111.90	140.40	(151.76)	(150.82)	(153.76)	**104.37**
bod_tibt	**118.58**	142.65	(134.47)	(137.26)	(206.06)	120.94
bos_latn	73.84	113.42	**63.67**	(155.27)	(135.77)	67.75
bug_latn	**172.63**	181.80	(206.57)	(214.07)	(218.78)	(211.21)
bul_cyrl	71.36	109.94	64.51	**59.03**	(148.72)	70.69
cat_latn	76.30	115.98	65.93	60.65	**60.55**	66.57
ceb_latn	**94.04**	125.05	(111.22)	(192.82)	(171.75)	103.37
ces_latn	75.07	115.11	**63.68**	(154.38)	(140.67)	71.35
cjk_latn	219.90	239.42	(212.05)	(211.78)	(220.85)	**189.97**
ckb_arab	**89.11**	121.96	(112.62)	(290.64)	(199.22)	107.81
crh_latn	**105.56**	127.10	(175.84)	(185.40)	(180.14)	119.20
cym_latn	**77.50**	114.42	97.27	(226.16)	(204.15)	100.22
dan_latn	74.09	111.50	**60.26**	(122.87)	(129.58)	67.08
deu_latn	73.91	118.06	57.93	**57.08**	(98.55)	62.85
dik_latn	**152.14**	169.32	(197.55)	(202.82)	(216.99)	(225.78)
dyu_latn	**183.05**	210.69	(209.72)	(216.63)	(209.24)	189.79
dzo_tibt	125.44	144.63	(213.72)	(217.28)	(238.66)	**110.58**
ell_grek	78.38	119.29	68.55	**65.99**	(157.23)	93.41
eng_latn	68.73	103.16	51.10	50.39	50.56	**48.43**
epo_latn	**75.57**	110.48	118.68	(175.58)	(149.97)	88.92
est_latn	73.18	109.48	71.72	**60.94**	(173.50)	96.32
eus_latn	70.76	103.94	112.97	**70.24**	70.78	101.32
ewe_latn	**128.71**	144.86	(181.68)	(202.81)	(228.49)	137.60
fao_latn	**89.22**	120.77	(182.64)	(216.89)	(180.74)	100.02
fij_latn	**107.24**	126.01	(186.08)	(136.05)	(215.93)	(121.46)
fin_latn	75.34	114.20	63.80	**55.50**	(164.25)	82.10
fon_latn	190.75	205.52	(209.87)	(255.39)	239.02	**190.66**
fra_latn	70.55	116.07	56.67	55.76	**53.43**	59.64
fur_latn	114.09	131.26	(203.28)	(198.54)	(185.86)	**107.32**
fuv_latn	**165.36**	188.23	(188.00)	(201.33)	(192.94)	(184.41)
gaz_latn	**120.21**	202.78	(163.96)	(260.02)	(262.77)	(144.71)
gla_latn	**97.03**	133.97	172.79	(245.08)	(209.10)	120.94
gle_latn	**79.48**	118.25	150.60	(228.43)	(204.19)	106.38

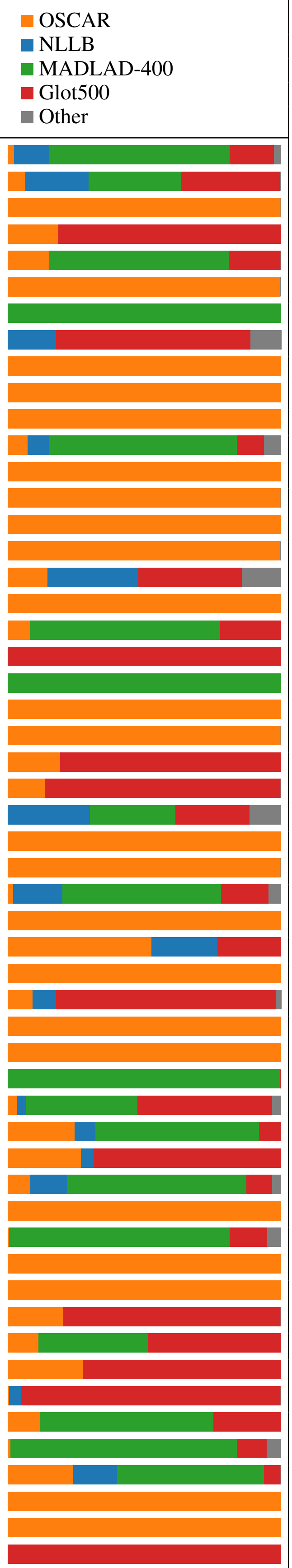
glg_latn	76.34	114.53	86.26	(107.84)	(87.42)	**74.13**
grn_latn	121.55	141.64	193.46	(230.52)	(225.65)	**119.96**
guj_gujr	**84.11**	110.95	100.35	(276.27)	96.00	97.04
hat_latn	89.82	114.98	114.15	**85.51**	(163.36)	102.63
hau_latn	**86.28**	116.50	115.81	(215.68)	(215.07)	107.27
heb_hebr	77.17	108.77	**73.71**	(175.48)	(147.78)	111.24
hin_deva	76.64	110.74	79.74	71.13	**70.06**	88.97
hne_deva	141.43	149.99	(164.54)	(153.61)	(144.13)	**110.53**
hrv_latn	71.16	110.07	**61.93**	(153.35)	(135.73)	66.55
hun_latn	75.05	113.40	**64.13**	(176.50)	(164.44)	78.54
hye_armn	**78.37**	115.07	85.23	(259.42)	(224.51)	94.83
ibo_latn	**110.84**	135.25	147.33	(248.34)	149.46	123.89
ilo_latn	**111.06**	129.57	(133.69)	(227.74)	(209.68)	115.75
ind_latn	72.11	101.11	62.46	60.13	**59.44**	65.23
isl_latn	**75.36**	113.10	82.56	(200.00)	(179.47)	93.17
ita_latn	75.27	116.41	60.06	**59.00**	(86.91)	62.98
jav_latn	**90.19**	112.98	102.58	(161.56)	(154.48)	102.21
jpn_jpan	68.51	100.99	63.11	**61.53**	(93.28)	68.24
kab_latn	**134.05**	154.27	(214.03)	(256.69)	(215.19)	159.99
kac_latn	**128.22**	137.82	(187.08)	(248.17)	(239.55)	146.59
kam_latn	(240.45)	(264.47)	(186.89)	(202.02)	(200.47)	**172.19**
kan_knda	**76.34**	103.10	92.36	(268.95)	93.76	97.03
kas_arab	(252.81)	(304.08)	(276.83)	(267.41)	**(245.94)**	(260.94)
kas_deva	**221.04**	240.29	(235.98)	(228.86)	(231.57)	(246.51)
kat_geor	**72.44**	108.39	82.24	(261.79)	(236.64)	96.53
kaz_cyrl	**73.52**	102.69	76.14	(192.47)	(186.61)	89.71
kbp_latn	145.05	159.84	(217.85)	(264.27)	(286.31)	**143.65**
kea_latn	145.29	157.19	(180.88)	(187.08)	(182.18)	**135.49**
khk_cyrl	**77.40**	103.60	87.49	(238.21)	(192.81)	97.14
khm_khmr	**98.82**	139.14	114.88	(323.85)	(240.85)	117.47
kik_latn	165.85	177.82	(222.18)	(244.09)	227.12	**148.48**
kin_latn	**87.00**	118.79	(227.82)	(225.85)	135.94	113.52
kir_cyrl	**70.91**	99.21	114.55	(224.69)	(202.97)	94.96
kmb_latn	179.18	199.36	(205.24)	(196.76)	(216.45)	**168.66**
kmr_latn	**99.93**	130.74	155.62	(234.52)	(213.35)	120.91
knc_arab	**181.38**	274.16	(223.08)	(222.78)	(214.43)	(228.52)
knc_latn	**170.17**	206.34	(229.82)	(227.51)	(239.24)	(242.18)
kon_latn	132.91	143.02	182.94	(190.80)	(189.74)	**(126.76)**
kor_hang	72.23	102.86	69.83	**63.54**	(122.61)	73.28
lao_laoo	**91.00**	120.61	110.10	(268.95)	(231.49)	107.59
lij_latn	140.38	168.14	(198.93)	(195.96)	(193.23)	**122.99**
lim_latn	123.42	154.99	(180.85)	(201.88)	(192.90)	**(116.94)**
lin_latn	**106.45**	122.73	144.57	(183.46)	167.57	116.17
lit_latn	71.55	110.06	**67.23**	(188.51)	(163.54)	92.94
lmo_latn	162.18	201.95	(203.26)	(199.20)	(198.36)	**152.10**
ltg_latn	**121.88**	138.23	(217.38)	(241.02)	(229.77)	(227.25)
ltz_latn	**85.00**	123.40	(154.55)	(149.83)	(188.39)	103.80
lua_latn	152.56	162.72	(175.04)	(167.40)	(201.83)	**147.08**
lug_latn	**118.73**	139.82	168.52	(213.70)	165.48	141.39
luo_latn	**139.03**	155.32	(210.18)	(218.56)	(222.85)	163.70
lus_latn	**95.13**	126.88	(157.04)	(217.70)	(211.72)	131.48
lvs_latn	70.94	110.34	**70.45**	(187.39)	(179.57)	86.42
mag_deva	**126.54**	139.42	(156.88)	(143.63)	(128.48)	(133.86)
mai_deva	123.50	142.56	(174.69)	(165.19)	(126.39)	**102.51**
mal_mlym	**80.56**	111.47	92.78	(221.87)	88.15	99.60
mar_deva	**82.32**	110.97	94.21	(200.05)	92.70	100.33
min_arab	(308.31)	(399.00)	(269.74)	(273.93)	**(254.70)**	(275.42)
min_latn	**108.38**	128.37	(167.37)	(164.01)	(160.32)	125.86
mkd_cyrl	**73.08**	109.43	73.22	(134.89)	(162.92)	76.94
mlt_latn	**83.60**	125.90	(280.75)	(279.00)	(237.59)	90.70
mni_beng	**(176.58)**	(274.13)	(279.36)	(271.00)	(188.18)	(275.78)
mos_latn	**187.64**	198.01	(228.35)	(236.65)	(241.18)	188.11
mri_latn	**97.39**	130.22	(191.98)	(187.99)	(180.89)	109.67
mya_mymr	**86.45**	125.64	119.52	90.49	(224.44)	121.00
nld_latn	71.72	112.22	**60.19**	(111.54)	(115.43)	65.13
nno_latn	80.80	114.77	**(73.23)**	(153.12)	(151.17)	76.16
nob_latn	76.13	109.96	**64.70**	(122.09)	(131.71)	68.04
npi_deva	**82.42**	111.11	90.72	(193.40)	86.21	86.53
nso_latn	**119.03**	139.17	(166.19)	(234.76)	181.13	123.81

nus_latn	(**217.98**)	(276.12)	(259.51)	(266.18)	(293.48)	(319.71)
nya_latn	**97.55**	121.06	(213.77)	(202.58)	180.39	118.77
oci_latn	97.75	135.49	(152.88)	(126.52)	(129.14)	**93.79**
ory_orya	**87.64**	114.38	(113.19)	(387.80)	(103.65)	105.48
pag_latn	122.99	139.15	(175.08)	(184.33)	(184.39)	**120.19**
pan_guru	**85.22**	116.57	107.69	(298.41)	99.43	101.39
pap_latn	**94.63**	126.00	(166.14)	(177.79)	(183.93)	118.24
pbt_arab	**104.87**	136.61	120.56	(263.66)	(210.05)	130.62
pes_arab	75.11	112.64	**70.67**	(152.72)	(145.67)	84.23
plt_latn	**87.89**	121.62	116.37	(236.48)	(176.35)	102.04
pol_latn	72.87	114.38	**61.18**	(146.68)	(130.92)	69.19
por_latn	72.38	110.34	59.80	57.79	**55.73**	63.06
prs_arab	96.31	120.44	(**80.74**)	(150.59)	(141.22)	83.70
quy_latn	**121.48**	144.34	185.50	125.42	(196.23)	132.70
ron_latn	75.68	118.45	**62.78**	(151.24)	(135.97)	70.39
run_latn	**112.80**	135.64	(229.50)	(226.86)	152.96	127.91
rus_cyrl	73.59	117.13	58.22	**57.38**	(110.95)	65.18
sag_latn	162.70	167.88	(182.49)	(171.77)	(196.64)	**150.49**
san_deva	**134.36**	156.91	167.12	(188.71)	(167.33)	140.17
sat_olck	148.03	156.26	(217.91)	(217.35)	(298.99)	**124.11**
scn_latn	124.37	157.13	(172.60)	(187.74)	(175.53)	**108.90**
shn_mymr	**162.52**	182.33	(253.11)	(475.53)	(373.38)	(341.88)
sin_sinh	**83.19**	114.20	97.91	(304.98)	(235.47)	106.82
slk_latn	72.22	113.10	**63.60**	(181.84)	(157.25)	78.82
slv_latn	71.40	111.05	**64.99**	(176.07)	(147.81)	75.62
smo_latn	**103.20**	142.22	(197.35)	(213.38)	(214.55)	120.15
sna_latn	**93.53**	123.08	(228.56)	(225.74)	180.82	125.19
snd_arab	**91.04**	116.69	150.37	(257.35)	(212.74)	117.36
som_latn	**104.72**	138.92	125.40	(245.50)	(229.18)	135.24
sot_latn	**99.85**	142.34	(171.86)	(235.46)	192.76	122.30
spa_latn	77.45	116.81	63.10	61.91	**59.82**	66.71
srd_latn	116.75	135.26	(202.61)	(200.61)	(191.63)	**106.39**
srp_cyrl	76.66	116.32	72.33	(175.02)	(149.07)	**71.86**
ssw_latn	**124.51**	143.31	186.61	(232.90)	(220.26)	133.78
sun_latn	**91.07**	116.80	109.40	(167.76)	(162.37)	98.90
swe_latn	73.76	109.95	**62.27**	(106.34)	(126.73)	65.90
swh_latn	79.45	109.31	89.60	**76.81**	98.22	95.63
szl_latn	131.79	157.70	(184.61)	(229.41)	(208.14)	**122.68**
tam_taml	79.87	107.64	88.40	**78.39**	79.09	95.70
taq_latn	(241.82)	(267.82)	(**216.97**)	(231.56)	(222.07)	(225.98)
taq_tfng	None	None	(301.75)	(289.98)	(261.59)	(396.24)
tat_cyrl	**76.67**	107.20	(222.35)	(227.31)	(192.83)	88.82
tel_telu	80.76	106.79	89.99	**77.70**	88.52	95.52
tgk_cyrl	**79.61**	113.99	(267.56)	(281.19)	(213.45)	100.89
tgl_latn	**86.11**	123.04	91.13	(168.86)	(162.76)	89.23
tha_thai	73.84	102.82	71.03	**64.31**	(177.26)	87.04
tir_ethi	**107.10**	134.77	142.22	(300.08)	(258.62)	133.08
tpi_latn	141.91	164.61	(218.27)	(212.39)	(197.37)	**109.12**
tsn_latn	**111.75**	143.35	184.89	(240.07)	186.95	127.47
tso_latn	**111.93**	133.30	(236.13)	(239.86)	205.36	130.50
tuk_latn	**81.23**	107.98	(239.28)	(256.07)	(198.90)	114.93
tum_latn	**132.63**	154.17	(233.00)	(187.16)	237.42	141.81
tur_latn	69.75	100.52	64.44	**61.13**	(130.11)	86.26
twi_latn	131.51	148.42	(211.42)	(216.89)	174.63	**128.30**
tzm_tfng	None	None	(206.77)	(206.42)	(243.59)	(332.15)
uig_arab	**75.68**	105.62	(317.09)	(367.46)	(206.46)	105.21
ukr_cyrl	76.60	116.07	**64.77**	(129.73)	(145.72)	72.19
umb_latn	182.34	211.34	(199.72)	(209.91)	(221.59)	**174.09**
urd_arab	83.96	117.74	90.05	**79.04**	85.58	98.80
uzn_latn	**71.09**	107.26	112.09	(243.88)	(217.18)	96.67
vec_latn	114.88	147.99	(161.71)	(155.44)	(160.51)	**108.87**
vie_latn	77.66	120.08	68.06	64.89	**61.00**	74.36
war_latn	**118.02**	153.85	(161.42)	(203.73)	(174.18)	132.17
wol_latn	**141.12**	158.76	202.72	(225.23)	167.95	161.96
xho_latn	**93.68**	121.39	144.06	(216.55)	155.76	122.42
ydd_hebr	**109.90**	144.63	(260.53)	(286.49)	(210.28)	128.80
yor_latn	**123.24**	167.30	174.33	(246.05)	154.45	148.67
yue_hant	90.03	121.49	(70.31)	(86.41)	**61.81**	69.76
zho_hans	78.92	121.82	66.34	65.42	**59.08**	70.09

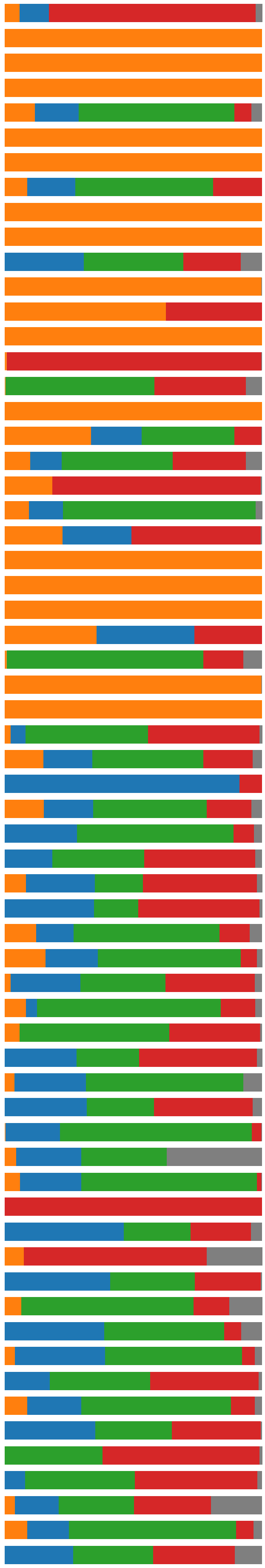
zho_hant	93.56	125.41	(72.53)	(89.36)	**63.06**	(75.40)
zsm_latn	73.25	100.45	**67.03**	(83.84)	(76.22)	72.90
zul_latn	**87.90**	118.40	135.47	(218.33)	186.47	115.47

Table 6: Goldfish languages with corresponding dataset sizes.

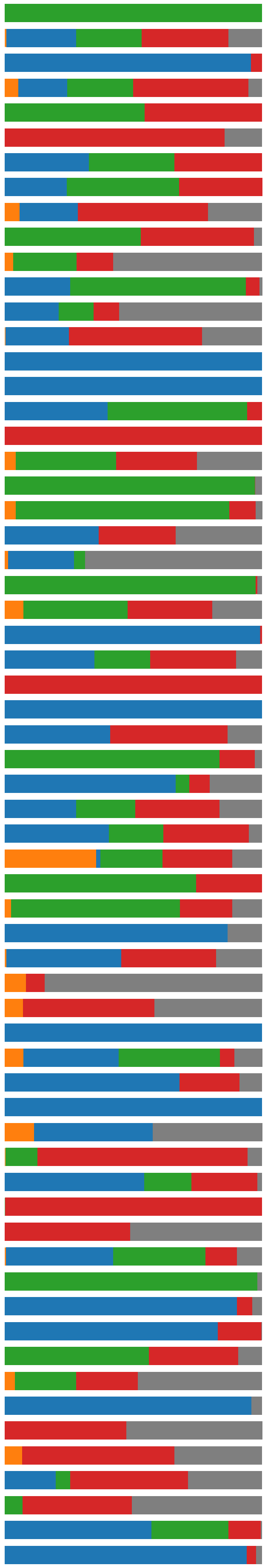
Language	Language (ISO 639-3)	Script (ISO 15924)	Byte Premium	Scaled MB	Tokens	Dataset Proportions
Afrikaans	afr	latn	1.04	1000.00	239682048	
Amharic	amh	ethi	1.72	1000.00	211767808	
Standard Arabic	arb	arab	1.47	1000.00	196197376	
Azerbaijani	aze	latn	1.30	1000.00	233091584	
Belarusian	bel	cyrl	2.01	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	



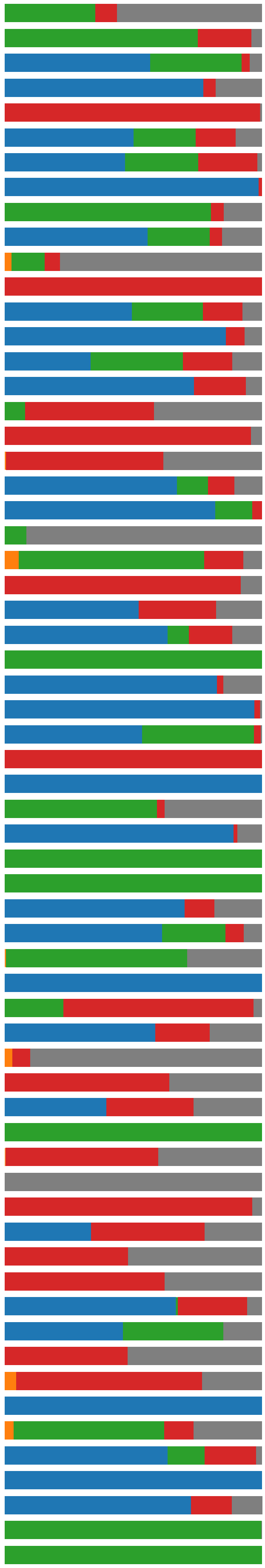
Panjabi	pan	guru	2.22	1000.00	215775232
Iranian Persian	pes	arab	1.60	1000.00	215946240
Polish	pol	latn	1.08	1000.00	216235008
Portuguese	por	latn	1.10	1000.00	225242112
Pushto	pus	arab	1.59	1000.00	237871616
Romanian	ron	latn	1.12	1000.00	230580224
Russian	rus	cyrl	1.82	1000.00	220467712
Sinhala	sin	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



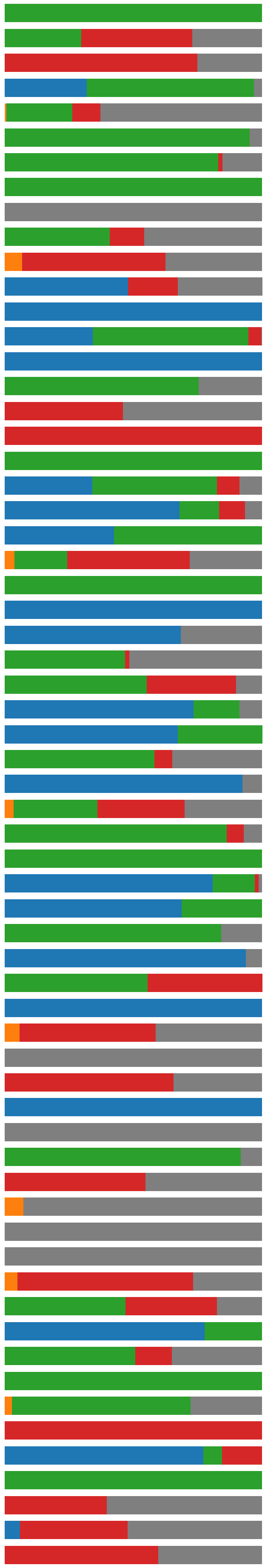
Modern Greek	ell	latn	1.24	376.42	92225536
Occitan	oci	latn	1.01	375.38	99783680
Plateau Malagasy	plt	latn	1.15	370.58	97517568
Assamese	asm	beng	2.53	348.88	77216256
Hmong	hmn	latn	1.19	345.97	100051968
Tosk Albanian	als	latn	1.17	336.30	87609344
Southern Sotho	sot	latn	1.17	332.91	94144000
Samoan	smo	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