Fixing the Foundation: Principled Approaches to Data Curation and Scaling

Sneha Kudugunta

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:
Luke Zettlemoyer, Chair
Pang Wei Koh
Ranjay Krishna

Program Authorized to Offer Degree:
Computer Science & Engineering

University of Washington

Abstract

Fixing the Foundation: Principled Approaches to Data Curation and Scaling

Sneha Kudugunta

Chair of the Supervisory Committee:
Professor Luke Zettlemoyer
Computer Science & Engineering

Training models at the scales used today (Comanici et al., 2025; OpenAI, 2025; Meta, 2025) is dependent on (1) curating large-scale datasets and (2) perfecting how we make scaling decisions. The datasets collected for pretraining contain trillions of tokens across different modalities, domains, and languages, often from messy sources such as crawled internet content. This necessitates extensive curation, filtering, and cleaning. Then, practitioners must determine the optimal training setting for their desired compute budget—typically by extrapolating from smaller training runs using scaling laws. Yet these two pillars are often studied separately: modern models are trained on heterogeneous and actively curated datasets like `MADLAD-400`, but most scaling-law research assumes a fixed data distribution and does not account for how changing dataset composition affects scaling behavior.

In this dissertation, I present principled approaches to these interconnected challenges. First, I discuss `MADLAD-400`, a manually audited, general-domain 3T token monolingual dataset based on CommonCrawl, spanning 419 languages. By documenting the curation process, I demonstrate how iterative audits surface failure modes that automatic filters miss. Second, I conduct a meta-analysis of more than 50 recent papers on scaling laws in deep learning, and identify key methodological decisions involved in scaling law research. I use this to explain discrepancies in the conclusions that several prior works reach, and empirically explore how significantly conclusions can vary depending on experimental design choices.

Third, I present ATLAS (Adaptive Transfer Scaling Laws), the largest multilingual scaling laws study to date — 774 training experiments spanning 10M–8B parameters and 400+ languages. I derive a language-agnostic scaling law for how specific languages in `MADLAD-400` scale with added compute that improves out-of-sample R^2 by more than 0.3 over prior work, and identify compute regimes in which practitioners should pretrain from scratch versus finetune from multilingual checkpoints.

Through these contributions I show that progress in foundation models depends not only on scaling up but on making the ingredients and assumptions of scaling explicit. Together, this work offers tools and empirical guidance for building more inclusive and compute-efficient foundation models.

ACKNOWLEDGMENTS

I am deeply grateful to the many people who made this dissertation possible.

I want to begin by thanking my advisor, Luke Zettlemoyer, for his guidance and support throughout what has been a somewhat unconventional PhD. I am thankful for his advice throughout the years, his patience through my rants every time I was frustrated, and his steady advice each time I was lost. I am especially grateful for the freedom Luke gave me to work on both the exciting directions described in this dissertation, and to carry several of the techniques I created during my PhD into model development at Gemini improving everything from data to architecture to post-training.

I also want to thank my PhD committee: Pang Wei Koh, Ranjay Krishna, and Lucy Lu Wang, for their support and interesting questions every time we met. Some of these questions pushed me to carry my proposed work over the finish line at moments when I struggled to believe in my own ideas. I would also like to thank the Kildall family for funding part of this journey.

I owe so much to my collaborators and mentors across academia and industry. I am glad to have learned and grown alongside this community, and to have pushed myself to become a better researcher and collaborator through working with all of you.

I am fortunate to have worked with Margaret, Danielle, Shayne, Niklas, Isaac, Sayna, Akari, Velocity, Terra, Hila, Yulia, Hanna, Biao, Xavier, Derrick, Aditya, Ankur, Orhan, Romi, Devvrit, Tim, Kaifeng, Inderjit, Sham, Ali, Prateek, Machel, Sebastian and many more. All the amazing research we've published together would not have been possible without you.

I will dearly miss my weekly lunches with ZLab: Ananya, Arti, Jacqueline, Joel, Hila, Kalyani, Luiza, Margaret, Mikel, Rulin, Sewon, Sahil, Suchin, Terra, Tim, Tong and Weijia. Thank you for teaching me so much each week, and for patiently listening to my latest rabbit holes. Thank you for

the sprawling discussions after lunch about life, research, purpose and everything under the sun that grew into great friendships and fruitful collaborations.

My manager at DeepMind, Orhan Firat, and Prateek Jain have provided invaluable mentorship and support throughout my PhD, and I am lucky to have had them in my corner. I am grateful to the many colleagues who worked alongside me in the trenches of every Gemini sprint: Abhanshu, Anirudh, Anmol, Antoine, Avery, Geoff, Isaac, Jeevesh, Kanav, Kevin L, Kevin R, Lars, Marcella, Miteyan, Rakesh, Romi, Ulyana, Tomas, ZD and so many more amazing researchers and engineers. I am grateful for the support on these endeavors from Oriol Vinyals, Rohan Anil and Jeff Dean.

To my friends: thank you for the love and laughter that sustained me through this journey. Some of you became collaborators. Some of you dragged me away from my laptop onto the ski slopes and hiking trails and running paths to fill my cup. Thank you for believing in me before I fully believed in myself, and for showing up with khichdi and hugs when I needed it most. I have been lucky to have decades-old friendships that have flourished as we moved across continents to chase our dreams, and to make new friends who so generously made me feel at home in a new city. Thank you for being my village across both Bay Area and Seattle: Aisha, Ayushi, Shruti, Margaret, Kartik, Aditya, Sanjana, Sahil, Abhinav, Sukriti, Ankur, Prateek, Bhatia, Siddhant, Anand, Vineet, Bose, Aru and so many more, including friends I've made over the years from both the wider CS department and RCR.

Most importantly, I want to thank my family for forming my bedrock of support. Amma, Naana and Krithik: we have not always been on the same continent the past few years, yet your love has been enormous and steady. Whether it was dropping everything to come visit, pushing me to ask myself the hardest questions, or simply encouraging me to not take life that seriously - thank you for everything.

I come from three generations of eldest daughters who wanted to be scientists. I am the first to have earned my PhD, and I have my parents to thank for that. Thank you, Amma and Naana, for giving me the love and support I needed to be a researcher. I will always be grateful for the

way you showed up for me: driving me to every exam, knowing I was struggling before I even thought to ask, and making sacrifices so you could be there next to me when I needed you most. At times, you came up against the boundaries of what dutiful Indian parents were supposed to allow of their daughters. Each time, you dismantled those expectations to make space for me to grow. I feel lucky to have you, and I hope I can be one reason more parents feel brave enough to wholeheartedly support their daughters to grow up to be exactly who they want to be. Krithik, thank you for all the laughs and motivating me to be a healthier, more athletic person. Professional sports is so much harder than being an AI researcher. Being around you reminds me of how everything important in life is a long game you win with discipline and consistency. I hope to continue supporting you through every success, each bigger than the last.

Vihang, we have spent most of the past two years separated by states, and sometimes even oceans, but your love, kindness and support have been constant. Your affection and care have inspired me to become a kinder and braver person, and I couldn't be happier to be building this new chapter of life with you by my side.

Chapter 1

INTRODUCTION

1.1 Introduction

Foundation models (Comanici et al., 2025; OpenAI, 2025; Meta, 2025) are pretrained by collecting vast datasets from multiple diverse sources, then training models with billions, if not trillions, of parameters with carefully tuned hyperparameters. The performance of these models is inextricably linked to the quality of data and the optimality of the training setup: the quality and diversity of the training corpora has been shown to greatly influence downstream results (Longpre et al., 2023a), and so have hyperparameter choices such as batch size McCandlish et al. (2018); Goyal et al. (2024), learning rate (Wortsman et al., 2023), and the shape of the architecture (Kaplan et al., 2020; Clark et al., 2022).

Training models at the frontier requires collecting trillions of tokens worth of data. The majority of this naturally occurring data comes from scraping the web (Raffel et al., 2020). These scrapes are noisy and contain significant amounts of malformed content, repetitive boilerplate and undesirable content (Dodge et al., 2021), which necessitates filtering out undesirable data and correcting any issues (Longpre et al., 2023a). While data cleaning practices are widely documented for English, data cleaning practices for multilingual data are often scant. Moreover, applying a one-size-fits-all approach to the entire corpus may result in underrepresenting in-language data Caswell et al. (2020); Dodge et al. (2021), or underrepresenting minority groups (Dodge et al., 2021). Kreutzer et al. (2022) find that multilingual data may be mislabeled, low in diversity, low quality, or otherwise unusable. They advocate for manual audits of data that can be caught by even non-proficient speakers combined with automatic filters, and discuss the limitations of automatic analysis, particularly for low-resource languages. Multiple works on English centric datasets, too, caution against using similar filters on the entire corpus, and suggest using filters based on each subdomain (Dodge et al.,

2021; Longpre et al., 2023a).

In the first part of this report, we create a dataset MADLAD-400, a 419 language dataset across 3T tokens based on CommonCrawl. We demonstrate the importance of combining iterative manual audits with automatic filters that are refined with each round of audits by expanding the language coverage of this dataset to over 400, beyond the 100-200 languages that most CommonCrawl based multilingual corpora have. We discuss the limitations revealed by self-auditing MADLAD-400, and the role data auditing had in the dataset creation process.

We then train and release a 10.7B-parameter multilingual machine translation model on 250 billion tokens covering over 450 languages using publicly available data, and find that it is competitive with models that are significantly larger, and report the results on different domains. In addition, we train a 8B-parameter language model, and assess the results on few-shot translation.

However, data curation is only half of the equation when it comes to creating foundation models. Even with high-quality datasets, performance is bounded by how effectively resources are allocated during training. Practitioners must therefore determine the optimal training setting for their desired compute budget. Given the prohibitive cost of training large models, this is often done by extrapolating the optimal architecture and hyper parameters settings from smaller training runs by describing the relationship between, loss, or task performance, and scale. That is, researchers use "scaling laws" (Hestness et al., 2017; Kaplan et al., 2020). However, this methodology is far from standardized, and foundational studies (Kaplan et al., 2020; Hoffmann et al., 2022) have even reached contradictory conclusions on optimal scaling strategies. Subsequent work including our own (Porian et al., 2024; Besiroglu et al., 2024; Li et al., 2025b) has shown that these laws are incredibly fragile, with results varying dramatically based on seemingly minor, and often unreported, details in the experimental and curve-fitting setup. These decisions can result in significant underperformance of language models and hinder progress.

In the second part of this report, we survey over 50 papers and demonstrate how every component of the process of fitting scaling laws varies, from the specific equation being fit, to the training setup, to the optimization method. Each of these factors may affect the fitted law, and therefore, the conclusions of a given study. We discuss examples of this in the literature, and augment this

discussion with our own analysis of the critical impact that changes in specific details may effect in a scaling study. We then discuss the resulting altered conclusions. To mitigate the underreporting of crucial details needed to reproduce findings, we propose a checklist for authors to consider while contributing to scaling law research.

Most of the papers we consider study scaling laws under a fixed data distribution. However, the composition of the dataset can greatly affect how the final model performs (Longpre et al., 2023a). Moreover, datasets within a mixture interact differently with increasing scale. To this end, in Chapter 4 I present ATLAS (Adaptive Transfer Scaling Laws), the largest multilingual scaling laws study to date — 774 training experiments spanning 10M to 8B parameters, 400+ training languages, and 48 evaluation languages, all built on `MADLAD-400`. ATLAS introduces a scaling law form that explicitly models cross-lingual transfer between languages, improving out-of-sample R^2 by more than 0.3 over prior work. We consider how scaling exponents differ across languages, and which source languages help (or hurt) which target languages. Then, we study how model size and data scale together as language coverage grows (quantifying the "curse of multilinguality"). Finally, we study when it is more compute-efficient to pretrain from scratch versus finetune from a multilingual checkpoint (a 144B–283B token crossover, depending on language).

Together, these contributions argue for a more principled view of foundation model development. Data curation and scaling are often discussed separately: in practice, we show that they are inseparable. By studying both pillars together, this dissertation provides tools and empirical guidance for building more compute optimal foundation models.

Chapter 2

MADLAD-400: A MULTILINGUAL AND DOCUMENT-LEVEL LARGE AUDITED DATASET

2.1 MADLAD-400: A Multilingual And Document-Level Large Audited Dataset

2.1.1 Introduction

The availability of large multilingual corpora has accelerated the progress of multilingual natural language processing (NLP) models (Xue et al., 2020; Conneau et al., 2019; Lin et al., 2021; Bapna et al., 2022; NLLBTeam et al., 2022). However, most publicly available general-domain multilingual corpora contain 100-200 languages (Xue et al., 2020; NLLBTeam et al., 2022; Abadji et al., 2022), with some datasets containing more languages in specific domains such as religious content Agic & Vulic (2019), children's books (Leong et al., 2022) or dialects (Adebara et al., 2022).

A common approach to creating such datasets is to mine language specific data from general web crawls such as CommonCrawl (Raffel et al., 2020; Laurençon et al., 2022; Xue et al., 2020) to create datasets. We simply take this approach and scale it. We train a document-level LangID model on 498 languages to obtain CommonCrawl annotations at a document level and obtain a 5-trillion token, document-level monolingual dataset.

However, such web-scale corpora are known to be noisy and contain undesirable content (Paullada et al., 2021; Luccioni & Viviano, 2021; Dodge et al., 2021), with their multilingual partitions often having their own specific issues such as unusable text, misaligned and mislabeled/ambiguously labeled data (Kreutzer et al., 2022). To mitigate this, we manually audit our data. Based on our findings, we discard 79 of the languages from our preliminary dataset, rename or combine several languages and apply additional preprocessing steps. Finally, to validate the efficacy of our dataset, we train multilingual machine translation models of various sizes up to 10.7B parameters, as well as an 8B decoder-only model, and then evaluate these models on highly multilingual translation

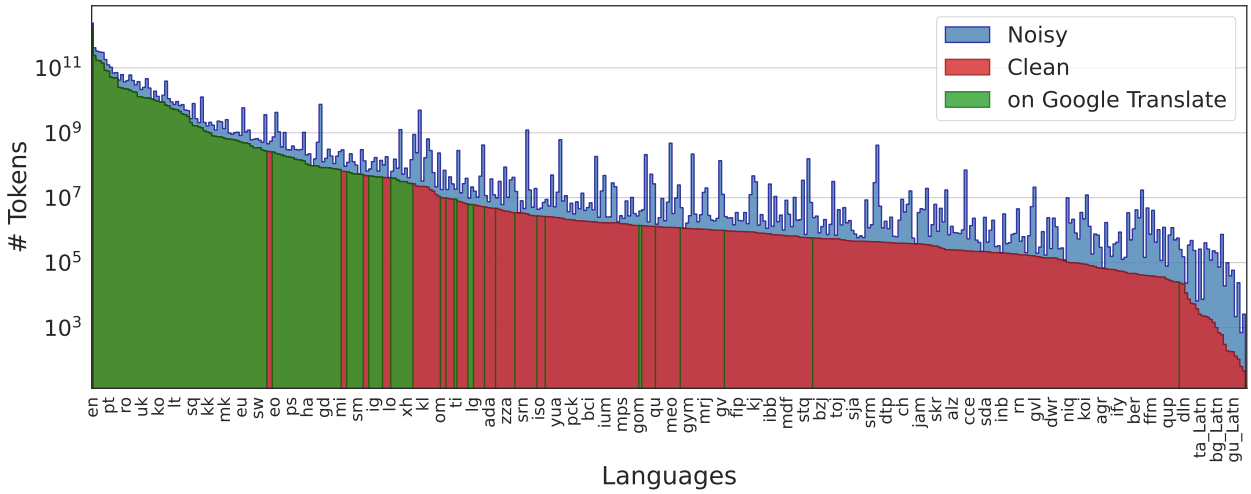


Figure 2.1: **Comparing the size of the noisy and clean monolingual datasets in `MADLAD-400`.** The difference is more noticeable on lower-resource languages, where noise effects are especially severe. For reference, languages supported by Google Translate are shaded in green. Note that, since this chart is in log scale, the difference in size is much greater than it may appear; for instance, for the lower-resource half of the dataset, the ratio is about 4× on median.

evaluation sets.

In Section 3.1.7, we describe the creation and composition of `MADLAD-400`, and discuss the results of the audit. Then, in Section 2.1.3, we describe the parallel data we collect using publicly available sources to train the multilingual machine translation models described in Section 2.1.4.1. In Section 2.1.4, we describe the training process of the multilingual machine translation models and 8B decoder-only model, and then evaluate these models on highly multilingual translation datasets. Finally, we discuss the limitations of this work and directions for future work.

2.1.2 `MADLAD-400`

The process we follow to create `MADLAD-400` is similar to that of other large-scale web corpora Caswell et al. (2020); Xue et al. (2020); Abadji et al. (2022); NLLBTeam et al. (2022). First, we collect as large a dataset of unlabeled web text as possible. More specifically, we use all available

snapshots of CommonCrawl[1] as of August 20, 2022. After some preliminary data cleaning, we use a highly multilingual LangID model to provide document-level annotations (Section 2.1.2.2). Finally, we conduct a self-audit (Section 2.1.2.4), or quality review, of this preliminary dataset partitioned by language, and design filters to remove noisy content. When appropriate, we correct language names and remove languages from the preliminary dataset. We note that building `MADLAD-400` was an iterative process, and that while we describe one major quality review in depth, we conducted several stages of filtering. To reflect this, we describe the preprocessing steps and improvements made in chronological order.

We release two version of this dataset: a 5 trillion token `noisy` dataset, which is the dataset obtained before applying document-level LangID and the final filters, and a 3 trillion token `clean` dataset, which has a variety of filters applied based on our self-audit, though it naturally has a fair amount of noise itself. Each dataset is released in both a document-level form and a sentence-level form. Some overall statistics for these dataset versions are given in Table 2.2, with a graph visualizing the distribution of sizes (number of tokens) across languages in Figure 2.1. The final version of `MADLAD-400` has 419 languages, with a varied geographic distribution, as seen in Table 2.1.

Table 2.1: Geographic distribution of languages in `MADLAD-400`.

Continent	# Languages
Asia	149
Americas	66
Africa	87
Europe	89
Oceania	26
Constructed	2

2.1.2.1 Preliminary Filters

We carry out a few preliminary preprocessing steps on the web-crawled corpus: first, we deduplicate lines across documents (Lee et al., 2021). Then, we filter out all pages that do not contain at least 3 lines of 200 or more characters (as done by Xue et al. (2020)). We also use other commonly used filtering heuristics such as removing lines containing the word "Javascript" and removing pages that

[1] https://commoncrawl.org/

Table 2.2: Overall statistics of both the noisy and clean partitions of MADLAD-400.

Dataset Version	# Documents		# Sentences		# Tokens	
	Total	Median	Total	Median	Total	Median
MADLAD-400-noisy	7.8B	27K	150B	240K	5.0T	7.1M
MADLAD-400-clean	4.0B	1.7K	100B	73K	2.8T	1.2M

contain "lorem ipsum" and curly brackets "{" (as done by Raffel et al. (2020)).

2.1.2.2 Language Identification (LangID)

We train a Semi-Supervised LangID model (SSLID) on 500 languages, following the recipe introduced by Caswell et al. (2020). We then filter the corpus on document-level LangID, which was taken to be the majority sentence-level LangID prediction. The resulting dataset is MADLAD-400-noisy. For the Additional details on these LangID models is in Appendix A.1.

2.1.2.3 Filtering Out Questionable Content

To assess the quality of this preliminary dataset, we inspected 20 sentences each from a subset of 30 languages in our dataset. Based on our observations, we introduced a score, pct_questionable. The pct_questionable score is simply the percentage of sentences in the input document that were "questionable". A sentence was considered questionable if any of the following were true:

1. **Document consistency:** Sentence-level LangID does not match the document-level LangID.
2. **List Case:** Over 50% percent of the tokens began in a capital letter (we apply this filter only if the sentence has at least 12 tokens.)
3. **Abnormal Lengths:** The sentence has under 20 characters or over 500 characters. We note that this is a bad heuristic for ideographic languages[2]).
4. **Technical Characters:** Over 20% of the characters in the sentence match [0-9{}+/()>].

[2] http://www.grcdi.nl/dqglossary/ideographic%20language.html

5. **Cursed Regexes:** The sentence matched a "cursed regex". These are a heuristic set of substrings and regexes that we found accounted for a significant amount of questionable content in the data samples we observed. They are described in depth in Appendix A.2.

We removed all documents with a `percent_questionable` score greater than 20%. Furthermore, we removed any document with under 5 sentences.

2.1.2.4 Self-Audit (Quality Review)

After filtering out generally lower-quality content with the approach described above, we performed a self-audit of every corpus in this dataset, following Kreutzer et al. (2022). The aim of our self-audit was to correct any remaining systematic issues by either applying additional filters, renaming/merging language codes, or completely removing the language from the dataset. Although we do not speak most of the 498 languages, we were able to give high-level comments on the general quality. For each language, we inspected a sample of 20 documents. This task was evenly divided between the first two authors based in part on which scripts they could read. We used the following guidelines:

- If dataset is mostly plausibly in-language text, we can keep it. For unknown languages, search the web for a few sentences and look at the website and URL for language clues.
- If dataset is noisy but the noise looks filterable, leave a note of how to filter it.
- If the dataset is very noisy and does not look possible to filter, mark it for removal.
- Optionally put note that may be helpful for downstream users, e.g. if dataset is 100% Bible.

We made the decision to include languages that looked noisy, but omit any language that was majority noise, or only had 20 or fewer docs. While this is not a high quality bar, we hope it still has the potential to be useful to the research community, given that foundation models have demonstrated the potential to learn distributions for very few examples Brown et al. (2020). The motivation for not releasing "nonsense" or tiny datasets is to avoid giving a false sense of how multilingual the dataset is ("Representation washing"), as recommended by **Quality at a Glance** (Kreutzer et al., 2022).

Overall Results. Of the 498 languages that we obtained LangID annotations for, we decided to omit 79 languages, bringing the final number of languages in `MADLAD-400` to 419. Based on the self-audit, we also expanded the filters (particularly the cursed regexes), and made changes as described in Sections 2.1.2.5 and 2.1.2.6. We details stats for these languages in Appendix Section A.4.

For transparency, we provide full results of the self-audit in Appendix A.4. In Table 2.3, we provide an overview of the issues surfaced through this self-audit. We find that a significant fraction of languages contain mostly or entirely religious documents, while other issues include misrendered text, pornographic content, and boilerplate.

Table 2.3: Summary of results of the audit on the preliminary dataset comprising of 498 languages. Note that there may be multiple issues with data in one language.

# Languages...	
Audited	498
With significant amounts of Bible data	141
With significant amounts of JW data	37
With significant amounts of LDS data	2
With significant amounts of virama-based issues	8
With a significant number of short docs	42
With complaints about noise	28
With complaints about porn	10
With complaints about boilerplate	15
With a note to remove from the dataset	77

2.1.2.5 Additional Filters

Based on the results of the self-audit, we apply three additional filters.

Virama Filtering and Correction. Many languages using Brahmic Abugida (South and South-east Asian scripts like Devanagari, Khmer, etc.) use some variant on the virama [3] character. We found that such languages in `MADLAD-400-noisy` had incorrectly encoded viramas: for example, तुम्हारे was rendered as तुम ् हारे, where the middle character is a detached virama. Therefore, for the languages `bn, my, pa, gu, or, ta, te, kn, ml, si, th, tl, mn, lo, bo, km, hi, mr, ne, gom, as, jv, dv, bho, dz, hne, ks_Deva, mag, mni, shn, yue, zh, ja, kjg, mnw, ksw, rki, mtr, mwr` and `xnr`, we did a special filtering/correction step — we removed all extraneous spaces before virama characters. We provide the pseudocode and list of virama characters in Appendix A.2.

Zawgyi Encoded Data. We found that languages using Myanmar script like `my` and `mnw` appeared to have the same issues with virama characters that still remained after applying the virama correction. This was because a large fraction of Myanmar script data on the internet is Zawgyi encoded data, which appears to have the rendering issues described above if rendered in Unicode. Therefore, we used an open-source Zawgyi detector [4] to convert the encoding of documents with more than a 50% probability of being Zawgyi encoded into standard Unicode encoding.

Chinese-Specific Filters. The Mandarin (`zh`) data in CommonCrawl had a particular issue with pornographic content. We combed through the data and developed a list of strings likely to be present in pornographic content, and filtered out all documents containing the strings in the blocklist. This resulted in a 17% reduction in the number of documents and a 56% reduction in file size. We list these strings in Appendix A.2.

---

[3] `https://en.wikipedia.org/wiki/Virama`

[4] `https://github.com/google/myanmar-tools`

2.1.2.6 Correcting Other Systematic Issues.

Based on various specific notes from the self-audit, we made a variety of changes. Five datasets were found to be in the wrong language, and were renamed or merged into the correct dataset. Six languages that looked suspicious were run by native speakers of those or related languages, some of which were discarded, and some of which were merged into the correct dataset. Finally, we removed all languages with fewer than 20 documents. Details can be seen in Appendix A.3.

2.1.3 Parallel Data

To train the machine translation (MT) models described in Section 2.1.4.1, we also collect a dataset composed of publicly available datasets coming from various data sources. A full list of the data sources and associated language pairs are in Appendix A.5. The final dataset has 156 languages across 4.1B sentence pairs and 4124 language pairs total. In the rest of the paper, we refer to the input sentence to an MT model as the “source side" and the reference/output sentence as the “target side".

2.1.3.1 Filters

We describe the data preprocessing steps taken below. We find that a significant amount of data is filtered out, with the amount of data available 396 of 4.1k language pairs reducing by more than 40%.

Deduplication. We deduplicate sentence pairs that are an exact match on both the source and target.

Virama Filtering and Correction/Zawgyi Encoded Data. We observed the same issues described in Section 2.1.2.5, and used the same filters for sentence pairs where either the source language or target language belonged to the list of languages in Section 2.1.2.5.

Unmatched Toxicity Filters. We use the unmatched toxicity filters described by NLLBTeam et al. (2022), but ultimately unusable for our purposes in most cases. For the languages ace, am, ar, az, bg, bm, bn, bs, cs, din, en, es, fa, fr, ga, gl, ha, hi, id, it, kk, ko, ml, ms, my, nl, no, nus, prs, ru, scn, sd, so, sv, tg, th, tt, ur, uz and zh, more than 3% of documents were marked as having unmatched toxicity. On closer inspection, we found that while zh and ko had a lot of pornographic content that was removed by the filtering process, most other languages removed sentences that had homonyms of non-toxic words. Similarly, languages like id, ur, tg, fa and no had data from Tanzil (Qur'an dataset), but the toxicity word lists contained words such as kafir, mercy and purity, that are not normally considered toxic content for our purpose of filtering the dataset using wordlists.

Source-Target Filters. We removed all sentences that have more than 75% overlap between the source and target side. To avoid filtering out valid entity translations, we only applied this filter on sentences longer than 5 tokens. In addition, we remove sentence pairs whose source length to target length ratio falls outside of $0.66 - 1.5$. We omitted this filter for the following, which are mainly non-whitespace languages: zh, ja, ko, km, my, lo, th, wuu, shn, zh_tw, zh_cn,iu, simple, dz, kr_Arab, din, nus and mi.

Script Filters. We removed all sentences that are less than 50% in-script for both the source and target language. For instance, if the sentence was supposed to be in kaa (Cyrillic script) but was 70% in the Latin script, we removed it.

2.1.3.2 Self-Audit (Quality Review)

Similar to the self-audit done for MADLAD-400, we conducted a review of the data sources that compose the parallel data we collected to verify the quality of this data. We collected 20 source-target pairs from each language, and assessed the data for the presence of offensive content, porn, and whether the data seemed to be of the correct language pair and whether the target sentence seemed to be a plausible translation. Since we did not have access to native speakers of all 157

languages, the latter was primarily based on guesses. In Appendix A.5 we provide full details of the instructions we provided to auditors, the results of the self-audit and any changes made the dataset.

2.1.3.3 A Note on Language Codes

As observed by Kreutzer et al. (2022), the datasets used to create the parallel data (and `MADLAD-400`) use a variety of different language codes. We use the BCP-47 standard, which specifies the 2-letter ISO-693-1 code when applicable, and otherwise the ISO-693-3 code. Script tags and region tags are omitted when they are defined as the default value by CLDR [5], and otherwise included. For example, `ks` refers to Kashmiri in Nastaliq/Arabic script (CLDR default), whereas `ks_Deva` refers to Kashmiri in Devanagari. A detailed investigation of codes in `MADLAD-400` can be found in Appendix A.3.

2.1.3.4 Multiway Data

We create additional multiway data by applying the n-gram matching method ($n = 8$) from Freitag & Firat (2020) to the processed dataset. Using this, and the publicly available 4.1k languages, we obtain 11.9B sentences across a total of 20742 language pairs. Full details may be found in Appendix A.7.

2.1.4 Experiments

We validate our data by training encoder-decoder machine translation models in Section 2.1.4.1 and decoder-only language models in Section 2.1.4.2, and test them on several translation benchmarks.

2.1.4.1 MT Models

We train models of various sizes: a 3B, 32-layer parameter model,[6] a 7.2B 48-layer parameter model and a 10.7B 32-layer parameter model. We share all parameters of the model across language pairs,

[5] `https://cldr.unicode.org/`

[6] Here and elsewhere, 'X-layer' means X encoder layers and also X decoder layers, for a total of $2X$ layers.

and use a Sentence Piece Model (Kudo, 2018) with 256k tokens shared on both the encoder and decoder side. Each input sentence has a <2xx> token prepended to the source sentence to indicate the target language (Johnson et al., 2017).

We use both supervised parallel data with a machine translation objective and the monolingual MADLAD-400 dataset with a MASS-style (Song et al., 2019) objective to train this model. Each of these objectives is sampled with a 50% probability. Within each task, we use the recently introduced UniMax (Chung et al., 2023) sampling strategy to sample languages from our imbalanced dataset with a threshold of $N = 10$ epochs for any particular language. We list specific architecture and training details of these models in Appendix A.8.

2.1.4.2 Zero-shot Translation with Language Models

Given recent interest in the efficacy of unsupervised translation using large language models, we explore training language models solely on the monolingual data. We follow the same training schedule and model configurations from Garcia et al. (2023). In particular, we consider 8B decoder-only models, following the same model hyperparameters as previous work (Chowdhery et al., 2022; Garcia et al., 2023). We train these models using a variant of the UL2 objective (Tay et al., 2022b) adapted for decoder-only models, and use the same configuration as previous work (Garcia et al., 2023; Orlanski et al., 2023). We provide additional details in Appendix A.8.

2.1.4.3 Evaluation

We use the sacreBLEU (Post, 2018) implementation of bleu[7] and chrf[8] as metrics. We evaluate our trained models on the following datasets:

WMT. We use the 15 WMT languages frequently used to evaluate multilingual machine translation models by Siddhant et al. (2022); Kim et al. (2021); Kudugunta et al. (2021); NLLBTeam et al.

[7] BLEU+case.mixed+lang.<sl>-<tl>+ numrefs.1+smooth.exp+tok.<tok>+version.1.3.0, tok=zh if tl=zh and 13a otherwise.

[8] nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1

(2022): cs, de, es, fi, fr, gu, hi, kk, lv, lt, ro, rs, es, tr and zh.

Flores-200. We evaluate on the languages in the Flores-200 dataset NLLBTeam et al. (2022) that overlap with the languages available in either MADLAD-400 or the parallel data described in Section 2.1.3. We list these languages in Appendix A.9. For non-English-centric pairs, we evaluate on a 272 language pair subset of the 40k language pairs possible due to computational constraints. We evaluate on all language pairs possible using the following languages as either source or target language: en, fr, cs, zh, et, mr, eu, cy, so, ckb, or, yo, ny, ti, ln, fon and ss. We obtained this set of languages by selecting every 10^{th} language by number of tokens in MADLAD-400 (clean), starting with French (fr). Noticing that this had no Indian languages, we shifted af and fo (both close dialects of HRLS) down one index to mr and or, respectively. Finally, we noticed that this initial list had supervised and unsupervised languages, but didn't have a good representative of a "slightly supervised language", that is, one with a small but extant amount of parallel data. Therefore, we added yo to the list, which has the least parallel data of any supervised language. This resulting subset of languages also contains a nice variety of scripts: Latin, Chinese, Devanagari, Arabic, Odia, and Ethiopic scripts.

NTREX. We evaluate on the languages in the recently introduced NTREX dataset (Federmann et al., 2022).

Gatones. Finally, we evaluate on the languages in GATONES, the in-house, 38-language eval set used in (Bapna et al., 2022) and the GATITOS paper (Jones et al., 2023). Again, we take the subset of languages overlapping with the languages available in either MADLAD-400 or the parallel training data.

Few-shot evaluation for language modeling We perform few-shot prompting to evaluate the language model with the following prompt:

[sl]:X_1\n[tl]:Y_1\n\n[sl]:X_2\n[tl]:Y_2\n\n...[sl]:X\n[tl]:

Table 2.4: Evaluation scores on WMT (depicted as <bleu> / <chrf>) for the MT models and language models described in Section 2.1.4.1 and Section 2.1.4.2 compared against NLLB-54B.

	NLLB	MT-3B	MT-7.2B	MT-10.7B	LM-8B			
					0-shot	1-shot	5-shot	10-shot
xx2en	34.2 / 60.4	33.4 / 60.0	34.9 / 60.6	**34.6 / 60.8**	2.3 / 17.3	25.1 / 51.4	26.2 / 52.9	26.2 / 53.4
en2xx	**31.1 / 58.0**	28.2 / 55.4	29.3 / 56.2	29.0 / 56.2	1.0 / 10.3	18.7 / 43.5	18.8 / 44.5	19.3 / 45.5
Average	**32.7 / 59.2**	30.8 / 57.7	32.1 / 58.4	31.8 / 58.5	1.6 / 13.8	21.9 / 47.4	22.5 / 48.7	22.8 / 49.4

Table 2.5: Evaluation scores on Flores-200 (depicted as <bleu> / <chrf>) for the MT models and language models described in Section 2.1.4.1 and Section 2.1.4.2 compared against NLLB-54B. All metrics are computed with the sacrebleu reference implementation.

	NLLB	MT-3B	MT-7.2B	MT-10.7B	LM-8B			
					0-shot	1-shot	5-shot	10-shot
xx2en	**35.5 / 59.6**	29.7 / 54.4	30.9 / 55.4	31.9 / 56.4	2.0 / 13.3	20.5 / 44.1	22.3 / 46.9	22.4 / 47.6
en2xx	**20.7 / 50.**1	17.3 / 44.1	17.8 / 44.7	18.6 / 45.7	0.4 / 5.7	8.1 / 26.7	8.7 / 29.0	8.7 / 28.8
Mean	**28.2 / 54.9**	23.5 / 49.2	24.4 / 50.0	25.3 / 51.1	1.2 / 9.6	14.3 / 35.5	15.6 / 38.0	15.6 / 38.2
xx2yy	**13.7 / 40.5**	8.8 / 31.2	8.4 / 30.9	10.1 / 34.0	0.3 / 4.1	4.0 / 16.1	4.4 / 17.3	4.2 / 17.1

where [sl] and [tl] denote the source and target language name (expressed in English. For example, when translating a sentence from en to te, we use [sl]=English and [tl]=Telugu), respectively. $X_\star$ and $Y_\star$ are demonstration examples used for prompting, and X is the test input.

For each test example, we randomly sample demonstration examples, which is simple yet performs competitively with more complicated strategies Vilar et al. (2022); Zhang et al. (2023). In particular, we randomly select examples from the dev split of each dataset. Since NTREX does not have a dev split, we randomly sample 1000 examples as the dev set and use the rest for test evaluation.

Table 2.6: Evaluation scores on the recently introduced NTREX test set (depicted as <bleu> / <chrf>) for the MT models and language models described in Section 2.1.4.1 and Section 2.1.4.2 compared against unsupervised baselines Baziotis et al. (2023). Note that LM-8B is evaluated on a 50% split of the NTREX data and is not comparable to the MT-model evaluations.

	Baziotis et al. (2023)	**MT-3B**	**MT-7.2B**	**MT-10.7B**	**LM-8B**			
					0-shot	1-shot	5-shot	10-shot
xx2en	- / -	30.6 / 54.5	32.7 / 56.2	**33.6 / 57.6**	3.2 / 17.3	20.4 / 43.8	23.8 / 48.2	24.4 / 49.0
en2xx	- / -	16.5 / 39.6	17.6 / **41.9**	**17.9 / 41.9**	0.8 / 7.3	11.7 / 31.2	12.6 / 32.4	12.3 / 32.3
Average	19.8 / -	23.5 / 47.0	25.1 / 49.0	**25.7 / 49.7**	2.0 / 12.3	16.0 / 37.4	18.1 / 40.2	18.3 / 40.6

2.1.4.4 Results

In Tables 2.4 and 2.6 we present evaluation scores on the WMT datasets and NTREX datasets, which are evaluation sets in the news domain. We find that both the 7.2B parameter model and the 10B parameter model is competitive with the significantly larger NLLB-54B model (NLLBTeam et al., 2022) on WMT. For the recent NTREX dataset, the only published results are small-scale results by Baziotis et al. (2023).

In Table 2.5 we find that on the Flores-200 dataset, our model is within 3.8 chrf of the 54B parameter NLLB model, while on **xxyy** pairs the 10.7B models is behind by 6.5 chrf. This is likely due to a combination of factors, including using a significantly smaller model (5x smaller), domain differences (Baziotis et al., 2023; Bapna et al., 2022), and back-translated data (Sennrich et al., 2016). Similarly, in Table 2.7, we find that the 10.7B parameter model is within 5.7 chrf of the scores reported by Bapna et al. (2022). Again, it is very difficult to compare their results to ours; their two largest advantages are 1) iterative back-translation, and 2) access to a much larger in-house text data. We note that the results we present are merely baselines to demonstrate the utility of MADLAD-400, and hope that future work builds upon these experiments by applying improved modeling techniques.

Across all evaluation datasets, we find that while results on few-shot translation using the

8B language model increase with an increasing number of demonstrations, these results are still significantly weaker than the results of models trained on supervised data. We present per-language pair results on all datasets in Appendix A.10.

Table 2.7: Evaluation scores on the GATONES test set used by Bapna et al. (2022) (depicted as <bleu> / <chrf>) for the MT models and language models described in Section 2.1.4.1 and Section 2.1.4.2.

	NTL (Bapna et al. (2022))		**MT-3B**	**MT-7.2B**	**MT-10.7B**	**LM-8B**			
	1.6B	6.4B				0-shot	1-shot	5-shot	10-shot
xx2en	- / 37.2	- / **41.2**	13.3 / 34.6	**15.5** / 36.5	14.8 / 36.0	0.3 / 6.5	6.6 / 25.4	8.3 / 28.1	8.4 / 28.4
en2xx	- / 28.5	- / **33.1**	4.5 / 23.9	**5.5** / 26.4	5.4 / 26.2	0.2 / 4.2	1.7 / 10.5	1.7 / 9.9	1.8 / 9.4
Average	- / 32.9	- / **37.2**	8.9 / 29.3	**10.5** / 31.5	10.1 / 31.1	0.3 / 5.4	4.2 / 18.0	5.0 / 19.0	5.1 / 18.9

2.1.5 Related Work

Extensive work has been done to mine general purpose datasets for multilingual machine translation and language modeling. Xue et al. (2020) introduce mC4, a general web domain corpus on 101 languages to train mT5, a pretrained language model for downstream NLP tasks. Similarly, Conneau et al. (2019) introduce CC-100, later extended to CC100-XL by Lin et al. (2021). The OSCAR corpus (Abadji et al., 2022) is also a mined dataset that supports 166 languages and the ROOTS corpus is a compiled dataset that contains 46 natural languages. Glot500-C (ImaniGooghari et al., 2023) covers 511 languages: however, it is not clear how many of these languages comprise solely of religious texts. Bapna et al. (2022) create an internal dataset on 1500+ languages, while NLLBTeam et al. (2022) mine a dataset from CommonCrawl and ParaCrawl (Esplà-Gomis et al., 2019). Recently, Leong et al. (2022) created a 350+ language dataset from children's books.

In addition, there have been efforts to get better represented corpora and models for languages often underrepresented in general multilingual corpora: Serengeti (Adebara et al., 2022) introduces

a dataset and associated model trained on 517 African languages and language varieties, while IndicTrans2 (Gala et al., 2023) introduces a machine translated model for the 22 scheduled languages in India.

2.1.6 Limitations

While we used thorough self-audits to guide the creation of MADLAD-400, we note that most audits were conducted by non-speakers of the languages in MADLAD-400; as a result, many types of noise, like machine-generated or disfluent content, could not be detected. Moreover, toxicity detectors, classifiers and filters that work reliably for all the 419 languages in MADLAD-400 do not exist, limiting the extent to which we can clean and document (Dodge et al., 2021; Bandy & Vincent, 2021) the dataset. It is possible that issues still remain, so we encourage users to report issues that will be listed on the project github page. This paucity extends to the availability of multilingual evaluation sets for these languages - we could only evaluate our models on 204 of the languages in MADLAD-400. Additionally, even though decoder-only models are often evaluated on NLP tasks that are not necessarily machine translation Hu et al. (2020); Asai et al. (2023); Ahuja et al. (2023), we did not conduct such evaluations - most available benchmarks cover only 30-50 languages of which most are not tail languages (which forms the focus of MADLAD-400). We instead leave this to future work. Finally, during our self-audit we noted the skew of data on the long tail towards specific domains such as religious texts. We hope that these limitations motivate the creation of more language-specific corpora not captured by web crawls, and the development of language-specific data cleaning tools and practices.

2.1.7 Conclusion

Through MADLAD-400, we introduce a highly multilingual, general web-domain, document-level text dataset. We perform a self-audit of this dataset for quality on samples of all 498 languages, develop filters, and remove spurious datasets, for a total of 419 languages in the release. We carefully describe the dataset creation process, laying out the iterations of audits and improvements upon the

preliminary dataset along with observations that guided our decisions. We hope that this encourages creators of large-scale pretraining datasets both to put in their due diligence for manually inspecting and dealing with data, and also to describe and publicize the process in a level of detail that is reproducible and insightful for downstream users. This increased visibility into the dataset creation cycle can in turn improve model development and enable responsible data use Sambasivan et al. (2021). Using `MADLAD-400`, we train and release large machine translation and general NLP models and evaluate them thoroughly. We hope that this further motivates work towards language technologies that are more inclusive of the rich language diversity housed by humanity.

Chapter 3

(MIS)FITTING SCALING LAWS: A SURVEY OF METHODOLOGICAL INCONSISTENCIES

3.1 (Mis)Fitting: A Survey of Scaling Laws

3.1.1 Introduction

While the previous section addressed data quality, training at the scale of recent large foundation models (Dubey et al., 2024; OpenAI, 2023; Reid et al., 2024) is an expensive and uncertain process. Given the infeasibility of hyperparameter tuning multi-billion parameter models, researchers extrapolate the optimal training setup from smaller training runs.

More precisely, scaling laws (Kaplan et al., 2020) are used to study many different aspects of model scaling. Scaling laws can guide targets for increasing dataset size and model size in pursuit of desired accuracy and latency for a specific deployment scenario, study architectural improvements, determine optimal hyperparameters and assist in model debugging. Scaling laws are often characterized as power laws between the loss and size of the model and dataset, and are seen in several variations (Section 3.1.2). These laws are found empirically by training models across a few orders of magnitude in model size and dataset size, and fitting the loss of these models to a proposed scaling law. Each component of this process varies in the reported literature, from the specific equation being fit, to the training setup, and the optimization method, as well as specific details for selecting checkpoints, counting parameters and the objective loss optimized during fitting.

Changes to this setup can lead to significant changes to the results, and therefore completely different conclusion to the study. For example, Kaplan et al. (2020) studied the optimal allocation of compute budget, and found that dataset size should be scaled more slowly than model size ($D \propto N^{0.74}$, D is dataset size, N is model size).

Later, Hoffmann et al. (2022) contradicted this finding, showing that model size and dataset size should be scaled roughly equally for optimal scaling. They highlight the differences in setup which lead to them showing that large models should be trained for significantly more tokens: particularly, they point to using later checkpoints, training larger models, a different learning rate schedule and changing the number of training tokens used across runs. Multiple followup works have focused on either reproducing or explaining the differences between these two papers (Besiroglu et al., 2024; Porian et al., 2024; Pearce & Song, 2024b). The authors find it challenging to reproduce results of previous papers - we refer the reader to Section 3.1.2 for a further discussion on these replication efforts.

Table 3.1: We provide a summary of the papers surveyed, highlighting the reproducibility challenges endemic to scaling law papers.

# Papers...	
Surveyed	51
With a quantified scaling law	45
With a description of training setup	40
With a definition of equation variables	36
With a description of evaluation	29
With a description of curve fitting	28
With analysis code	19
With metric scores or checkpoints provided	17

Motivated by this, we survey over 50 papers on scaling laws across a variety of modalities, tasks and architectures, and find that essential details needed to reproduce scaling law studies are often underreported. We broadly categorize these details as follows:

Section 3.1.3: What *form* are we fitting? Researchers may choose any number of power law forms relating any set of variables, to which they fit the data extracted from training runs. Even seemingly minor differences in form, may imply critical changes in assumptions – for example, about certain interactions between variables which are excluded, the definitions of these variables or error terms which are deemed significant enough to include.

Section 3.1.6: How do we *train models*? In order to fit a scaling law, one needs to train a range of models spanning orders of magnitude in parameter count and/or dataset size. Each model requires a

multitude of hyperparameter and parameter choices, such as the specific model/dataset sizes to use, the architecture shape, batch size or learning rate schedule.

Section 3.1.7: How do we *extract data* after training? Once these models are trained, downstream metrics like perplexity must be obtained from the intermediate or final checkpoints. This data may be also scaled, interpolated or bootstrapped to create more datapoints to fit the power law parameters.

Section 3.1.8: How are we *optimizing* the fit? Finally, the variable must be fit with an objective and optimization method, which may in turn have their own initialization and hyperparameters to choose.

To aid scaling laws researchers in reporting details necessary to reproduce their work, we propose a checklist (Figure 3.1 - an expanded version may be found in Appendix B.2). Based on this checklist, we summarize these details for all 51 papers in tabular form in Appendix B.3. We find that important details are frequently underreported, significantly impacting reproducibility, especially in cases where there is no code - only 19 of 42 papers surveyed have analysis code/code snippets available. Additionally, 23 (a little over half) of surveyed papers do not describe the optimization process, and 15 do not describe how training FLOPs or number of parameters are counted, which has been found to significantly change results (Porian et al., 2024). In addition, we fit our own power laws to further demonstrate how these choices critically impact the final scaling law (Section §3.1.9).

3.1.2 Papers on Scaling Laws

Researchers have proposed scaling laws to study the scaling of deep learning across multiple domains and for several tasks. Studies of the scaling properties of generalization error with training data size and model capacity predate modern deep learning. Banko & Brill (2001) observed a power law scaling of average validation error on a confusion set disambiguation task with increasing dataset size. The authors also claimed that the model size required to fit a given dataset grows log linearly. As for larger scale models, Amodei et al. (2016) observe a power-law WER improvement

Scaling Law Reproducilibility Checklist

Scaling Law Hypothesis (§3.1.3)
- Form
- Variables (Input)
- Parameters
- Derivation and Motivation
- Assumptions

Training Setup (§3.1.6)
- # of models
- Model size range
- Dataset source & size
- Parameter/ FLOP Count Calculation
- Hyperparameter Choice
- Other Settings
- Code Open Sourcing

Data Collection(§3.1.7)
- Checkpoint Open Sourcing
- # Checkpoints per Power Law
- Evaluation Dataset & Metric
- Metric Modification & Code

Fitting Algorithm (§3.1.8)
- Objective (Loss)
- Algorithm
- Optimization Hyperparameters
- Optimization Initialization
- Data Usage Coverage
- Validation of Law(s)

Figure 3.1: We introduce a checklist for researcher to use for scaling laws research. In Appendix B.2, we include an expanded version of the checklist that may be used as a template.

on increasing training data for a 38M parameter Deep Speech 2 model. Hestness et al. (2017) show similar power law relationships across several domains such as machine translation, language modeling, image processing and speech recognition. Moreover, they find that these exponential relationships found hold across model improvements.

Kaplan et al. (2020) push the scale of these studies further, studying power laws for models up to 1.5B parameters trained on 23B tokens to determine the optimal allocation of a fixed compute budget. Later studies (Hoffmann et al., 2022; Hu et al., 2024) revisit this and find that Kaplan et al. (2020) greatly underestimate the amount of data needed to train models optimally, though major procedural differences render it challenging to attribute the source of this discrepancy. Since then, researchers have studied various aspects of scaling up language models. Wei et al. (2022) examine the emergence of abilities with scale that are not present in smaller models, while Hernandez et al. (2021) study the scaling laws for transfer between distributions in a finetuning setting. Henighan et al. (2020) consider possible interactions between different modalities while recently, Aghajanyan et al. (2023) study scaling in multimodal foundation models. Tay et al. (2022a) show that not all architectures scale equally well, highlighting the importance of using scaling studies to guide architecture development. Poli et al. (2024) scale hybrid architectures like Mamba (Gu & Dao, 2023), showing the efficacy of this new model family. Other researchers also formulate specific scaling laws to study other Transformer based architectures. For example, Clark et al. (2022) and Frantar et al. (2023) introduce new scaling laws to study mixture of expert models (Fedus et al., 2022; Shazeer et al., 2017) and sparse models (Zhu & Gupta, 2017) respectively. Researchers have also used scaling laws to study encoder-decoder models for neural machine translation (Ghorbani et al., 2021; Gordon et al., 2021), and the effect of data quality and language on scaling coefficients (Bansal et al., 2022; Zhang et al., 2022). While language models form the majority of the papers surveyed, we also consider papers that study VLMs (Cherti et al., 2023; Henighan et al., 2020), vision (Alabdulmohsin et al., 2022; Zhai et al., 2022), reinforcement learning (Hilton et al., 2023; Jones, 2021; Gao et al., 2023) and recommendation systems (Ardalani et al., 2022). We further discuss different forms of scaling laws researchers introduce for the specific research questions they wish to answer in Section 3.1.3.

A majority of the surveyed papers study Transformer (Vaswani, 2017) based models, but a few consider different architectures. For example, Sorscher et al. (2022) investigate data pruning laws in ResNets, and some smaller scale studies use MLPs or SVMs (Hashimoto, 2021). This overrepresentation is perhaps partially a result of Transformer-based models achieving higher scale than other architectures; a ResNet101 has 44M parameters, while the largest Llama 3 model has 405B.

Replication Efforts Besiroglu et al. (2024) seek to reproduce the parameter fitting approach used by Hoffmann et al. (2022). They are unable to recover the scaling law from Hoffmann et al. (2022), and demonstrate that the claims of the original paper are inconsistent with descriptions of the setup. They then seek to improve the fit of the scaling law by initializing from the parameters found in Hoffmann et al. (2022) and modifying parts of the power law fitting process.

Porian et al. (2024) isolate several decisions as primarily responsible for the discrepancy between the recommendations of Kaplan et al. (2020) and Hoffmann et al. (2022): (1) learning rate scheduler warmup, (2) learning rate decay, (3) inclusion of certain parameters in total parameter count, and (4) specific training hyperparameters. By adjusting these factors, they are able to reproduce the results of Kaplan et al. (2020) and Hoffmann et al. (2022). However, they only use 16 training runs to fit their scaling laws, each designed to match one targeted setting (e.g., replicating Kaplan et al. (2020)). Instead of using raw loss values, they fit to loss values found by interpolating between checkpoints. Like Pearce & Song (2024b), they apply a log transform and linear regression to fit their law.

3.1.3 What form *are we fitting?*

A majority of papers we study fit some kind of power law ($f(x) = ax^{-k}$). That is, they specify an equation defining the relationship between multiple factors, such that a proportional change in one results in the proportional change of at least one other. They then optimize this power law to find some parameters. A few efforts do not seem to fit a power law, but may show a line of best fit, obtained through unspecified methods (Rae et al., 2021; Dettmers et al., 2022; Tay et al., 2022a;

Shin et al., 2023; Schaeffer et al., 2023; Poli et al., 2024).

The specific form may be motivated by researcher intuition, previous empirical results, prior work, code implementation, or data availability. More importantly, the form is often determined by the specific question(s) a paper investigates. For example, one may attempt to predict the performance achieved by scaling up different model architectures, or the optimal ratio for model scaling vs data scaling when increasing training compute (Kaplan et al., 2020; Hoffmann et al., 2022). Based on this, we loosely classify scaling laws by their form as *performance prediction* and *ratio optimization* approaches. We indicate this classification for all surveyed papers in Appendix B.3.

3.1.4 Ratio Optimization

The simplest scaling law forms usually predict the relation between two variables in an optimal setting. For example, approaches 1 and 2 from Hoffmann et al. (2022) fit to the optimal (i.e., lowest loss) D and N values for a particular compute budget C. Porian et al. (2024), aiming to resolve these inconsistencies, defines $\rho^* = \frac{D^*}{N^*}$ and writes this relationship as:

$$N^*(C) = N_0^* \cdot C^\alpha; D^*(C) = D_0^* \cdot C^\alpha; \rho^*(C) = \rho_0^* \cdot C^\alpha \tag{3.1}$$

They assume $C \approx 6ND$, and thus only need to fit the first equation; the other power laws can be inferred. This simplicity is deceptive in some cases, as collecting $(N^*(C), C)$ pairs may not be trivial. It is possible to fix C and follow a binary search approach to train a multitude of models, then bisect to approximate the performance-optimal N, D pair. However, this quickly grows prohibitively costly. In practice, it is common to interpolate between a set of fixed results to estimate the true $N^*(C)$ §3.1.7. This adds to the complexity of this approach, and introduces a hidden dependency on the performance evaluation, yet it does not actually predict the performance of the optimal points. If only the performance of the optimal-ratio model is of interest, it is possible to fit a second power law $L(N^*(C), D^*(C)) = a \cdot C^\alpha$. Most papers we survey choose to fit a power law which directly predicts performance.

3.1.5 Performance prediction

Kaplan et al. (2020) proposes a power law between Loss L, number of model Parameters N, and number of Dataset tokens D:

$$L(N, D) = \left[\left(\frac{N}{N_c} \right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D}{D_c} \right]^{\alpha_D} \tag{3.2}$$

On the other hand, Approach 3 of Hoffmann et al. (2022) proposes

$$L(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} \tag{3.3}$$

In both of the above, all variables other than L, N, and D are parameters to be found in the power law fitting process. Though these two forms are quite similar, they differ in some assumptions. Kaplan et al. (2020) constructs their form on the basis of 3 expected scaling law behaviors, and Hoffmann et al. (2022) explains in their Appendix D that their form is based on risk decomposition. The resulting Kaplan et al. (2020) form includes an interaction between N and D in order to satisfy a constraint requiring assymmetry introduced by one of their expected behaviors. The Hoffmann et al. (2022) form, on the other hand, consists of 3 additive sources of error, E representing the irreducible error that would exist even with infinite data and compute budget, as well as two terms representing the error introduced by limited parameters and limited data, respectively.

Power laws for performance prediction can sometimes yield closed form solutions for optimal ratios as well. However, the additional parameters and input variables, introduced by the need to incorporate the performance metric term, add random noise and dimensionality. This increases the difficulty of optimization convergence, so when prediction performance is not the aim, a ratio optimization approach is frequently a better choice.

Many papers directly adopt one of these forms, but some adapt these forms to study relationships with other input variables. Clark et al. (2022), for example, study routed Mixture-of-Expert models, and propose a scaling law that relates dense model size (effective parameters) N and number of experts E with a biquadratic interaction ($\log L(N, E) \triangleq a \log N + b \log E + c \log N \log E + d$). Frantar et al. (2023) study sparsified models, and propose a scaling law with an additional parameter sparsity S, the optimal value of which increases with N ($L(S, N, D) = \left(a_S(1 - S)^{b_S} + c_S\right) \cdot$

$\left(\frac{1}{N}\right)^{b_N} + \left(\frac{a_D}{D}\right)^{b_D} + c$). Other papers change the form to model variables in the data setup. Aghajanyan et al. (2023) consider interference and synergy between multiple data modalities ($L(N, D_j) = E_j + \frac{A_j}{N^{\alpha_j}} + \frac{B_j}{|D_j|^{\beta_j}}$, $L(N, D_i, D_j) = [\frac{L(N,D_i)+L(N,D_j)}{2}] - C_{i,j} + \frac{A_{i,j}}{N^{\alpha_{i,j}}} + \frac{B_{i,j}}{|D_i|+|D_j|^{\beta_{i,j}}}$), while Goyal et al. (2024), Fernandes et al. (2023) and Muennighoff et al. (2024) add terms to their scaling law formulations which represent mixing data sources and/or repeated data, using notions such as diminishing utility. A comprehensive list of the power law forms in the surveyed papers may be found in Table B.4.

3.1.6 How do we train models*?*

In order to fit a scaling law, one needs to train a range of models across multiple orders of magnitude in model size and/or dataset size. Researchers must first decide the range and distribution of N and D values for their training runs, in order to achieve stable convergence to a solution with high confidence, while limiting the total compute budget of all experiments. Many papers did not specify the number of data points used to fit each scaling law; those that did range from 4 to several hundred, but most used fewer than 50 data points. The specific N and D values also skew the optimization process towards a certain range of N/D ratios, which may be too narrow to include the true optimum. Some approaches, such as using IsoFLOPs (Hoffmann et al., 2022), additionally dictate rules for choosing N and D values. Moreover, using a minimum N or D value may result in outlier values that may need to be dropped (Porian et al., 2024; Shin et al., 2023; Henighan et al., 2020). We investigate this choice in Section §3.1.11

The definition of N, D, or compute cost C can affect the results of a scaling study. For example, if a study studies variation in tokenizers, a definition of training data size based on character count may be more appropriate than one based on token count (Tao et al., 2024). The inclusion or exclusion of embedding layer compute and parameters, may also skew the results of a study - a major factor in the different in optimal ratios determined by Kaplan et al. (2020) and Hoffmann et al. (2022) has been attributed to not factoring embedding FLOPs into the final compute cost (Pearce & Song, 2024b; Porian et al., 2024). Given the increase in extremely long context models (128k-1M) Reid et al. (2024), the commonly used training FLOPs approximation $C = 6ND$ (see Appendix

B.3) may not hold for such models, given the additional cost proportional to the context length and model dimension - Bi et al. (2024) introduce a new terms non-embedding FLOPs/token to account for this.

Scaling law fit depends on the performance of each individual checkpoint, which is highly dependent on factors such as training data source, architecture and hyperparameter choice. Bansal et al. (2022) and Goyal et al. (2024), for instance, discuss the effect of data quality and composition on power law exponents and constants. Repeating data has also been found to yield different scaling patterns in large language models (Muennighoff et al., 2024; Goyal et al., 2024).

Researchers have also studied the effect of architecture choice on scaling - Hestness et al. (2017) find that architectural improvements only shift the irreducible loss, while Poli et al. (2024) suggest that these improvements may be more significant. The way in which a model is scaled can also affect results. Within the same architecture family, Clark et al. (2022) show that increasing the number of experts in a routed language model has diminishing returns beyond a point, while Ghorbani et al. (2021) find that scaling the encoder and decoder have different effects on model performance. Scaling embedding size can also drastically change scaling trends (Tao et al., 2024).

The optimal hyperparameters to train a model changes with scale. Changing batch size, for example, can change model performance McCandlish et al. (2018); Kaplan et al. (2020). Optimal learning rate is another hyperparameter shown to change with scale, though techniques such as those proposed in Tensor Programs series of papers (Yang et al., 2022) can keep this factor constant with simple changes to initialization. More specifically, changing the learning rate schedule from a cosine decay to a constant learning rate with a cooldown (or even changing the learning rate hyperparameters) has been found to greatly affect the results of scaling laws studies (Hu et al., 2024; Porian et al., 2024; Hägele et al., 2024; Hoffmann et al., 2022).

One common motivation for fitting a scaling laws is extrapolation to higher compute budgets. However, there is no consensus on the orders of magnitude up that one can project a scaling law and still find it accurate, nor on the breadth of compute budgets that should be covered by the data. We find that the range of model size N and dataset size D greatly varies, with the maximum value of N in each paper ranging from 10M parameters to around 7B and that of D being as large as 400B

tokens. For most papers we survey, the scales are relatively modest: 13 of 51 papers train models beyond 2B parameters; most only train models smaller than 1B parameters. It has been shown, with some controversy Schaeffer et al. (2023), that scaling to significantly larger scales can result in new abilities that did not appear in smaller models (Wei et al., 2022). Forecasting limits to extrapolation and the appearance of new abilities at new scales is an open question.

3.1.7 How do we collect data *from model training?*

To evaluate the range of models trained to fit a scaling law, train or validation loss are most commonly used, but some works consider other metrics, such as ELO score (Jones, 2021; Neumann & Gros, 2022), reward model score (Gao et al., 2023), or downstream task metrics like accuracy or classification error rate (Henighan et al., 2020; Zhai et al., 2022; Cherti et al., 2023; Goyal et al., 2024; Gao et al., 2023). This choice is non-trivial - while some papers show that there is a power law relation between the predicted loss found by using validation loss and a different downstream task (Dubey et al., 2024), it is possible for the results of a study to change completely depending on the metric used. Schaeffer et al. (2023), for example, find that using linear metrics such as Token-edit distance instead of non-linear metrics such as accuracy produces smooth, continuous predictable changes in model performance, contrary to an earlier study by Wei et al. (2022). Moreover, Neumann & Gros (2022) find that they are unable to use test loss instead of Elo scores to fit a power law.

While it is most straightforward to evaluate only the final checkpoint on the target metric, some studies may use the median score of the last several checkpoints of each training run Ghorbani et al. (2021), or multiple intermediate checkpoints throughout each training run for various reasons. One common reason is that this is the only computationally feasible way to obtain a fit with sufficient confidence intervals (Besiroglu et al., 2024). For instance, the ISOFLop approach to finding the optimal D/N ratio in Hoffmann et al. (2022) requires training multiple models for each targeted FLOP budget - this would be computationally prohibitive to do without using intermediate checkpoints. Hoffmann et al. (2022), in particular, use the last 15% checkpoints. Some papers also report bootstrapping values (Ivgi et al., 2022). This detail is often not specified in scaling law papers, with only 29 of 51 papers reporting this information - we point the reader to Appendix B.3

for an overview.

A related technique is performance interpolation. Porian et al. (2024) do not aim to exactly match the desired FLOP counts when evaluating model checkpoints mid-training. They instead interpolate between multiple model checkpoints to estimate the performance of a model with the target number of FLOPs. Hoffmann et al. (2022) and Tao et al. (2024) also interpolate intermediate checkpoints. Hilton et al. (2023), relatedly, smooth the learning curve before extracting metric scores.

As discussed in Section 3.1.6, training models with too little data or too few parameters can skew the results. To prevent this issue, several works report filtering out data points before fitting their power law. Henighan et al. (2020) drop their smallest models, while Hilton et al. (2023) and Hoffmann et al. (2022) exclude early checkpoints. Muennighoff et al. (2024) remove outlier datapoints that perform badly due to excess parameters or excess epochs. Similarly, Ivgi et al. (2022) remove outlier solutions after bootstrapping.

3.1.8 How are we optimizing *the fit?*

The optimization of a power law requires several design decisions, including optimizer, loss, initialization values, and bootstrapping. We discuss each in this section. Over half of the papers we analyze do not provide any information about their power law fitting process, or provide limited information only and fail to detail crucial aspects. Specifically, many papers fail to describe their choice of optimizer or loss function. In Table 3.2, we provide an overview of the optimization details (if specified) for each paper considered.

Optimizer Power laws are most commonly fit with a variety of algorithms designed to optimize non-linear functions. One of the most common is the BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm, or a variation L-BFGS (Liu & Nocedal, 1989). Some papers (Hashimoto, 2021; Covert et al., 2024) use Adam, Adagrad, or other optimizers common in machine learning, such as AdamW, RMSProp, and SGD. Though effective for LLM training, these are sometimes ill-suited for the purpose of fitting a scaling law, due to various factors limiting their practicality, such as data-

hungriness. Goyal et al. (2024) forgo the use of an optimizer altogether due to instability of solutions (see initialization) and rely exclusively on grid search to fit their scaling law parameters.

Some scaling law works (Rosenfeld et al., 2019) opt to use a linear method, such as linear regression, which is generally much simpler. To do this, they typically convert the hypothesized power law to a linear form by taking the log. For example, for a power law $y^b = c \cdot x^a + d$, use the form $\beta \cdot log(y) = \gamma + \alpha \cdot log(x)$ instead. For loss prediction, this results in a form similar to $log(L(N, D)) = \alpha \cdot logN + \beta \cdot logD + E$. This trick is employed even when using an optimizer capable of operating on non-linear functions (Hashimoto, 2021). Though this conversion may sometimes work in practice, it is generally not advised because the log transformation also changes the distribution of errors, exaggerating the effects of errors at small values. This mismatch increases the likelihood of a poor fit (Goldstein et al., 2004). We found this approach to be very common among the papers we study.

Loss Various loss functions have been chosen for power law optimization, including variants on MAE (mean absolute error), MSE (mean squared error), and the Huber loss (Huber, 1992), which is identical to MSE for errors less than some value δ (a hyperparameter), but grows linearly, like MAE, for larger errors, effectively balancing the weighting of small errors with robustness to outliers. Of the papers which specify their loss function, most use a variant of MAE (Ghorbani et al., 2021), MSE (Goyal et al., 2024; Hilton et al., 2023), Huber loss (Hoffmann et al., 2022; Aghajanyan et al., 2023; Frantar et al., 2023; Tao et al., 2024; Muennighoff et al., 2024), or a custom loss (Covert et al., 2024).

Initialization Initialization can have a substantial impact on final optimization fit (§3.1.9). One approach is to iteratively train with different initializations, selecting the best fit at the termination of the search. This is typically a grid search over choices for each parameter (Aghajanyan et al., 2023; Muennighoff et al., 2024), or a random sample from that grid (Frantar et al., 2023; Tao et al., 2024). Alternatively, the full grid of potential initializations can be evaluated on the loss function without training, and the most optimal k used for initialization and optimization (Caballero et al.,

2022). Finally, if a hypothesis exists, either from prior work or expert knowledge about the function, this hypothesis may be used instead of a search, or to guide the search (Besiroglu et al., 2024).

Validating the Scaling Law A majority of the papers surveyed do not report validating the scaling law in any meaningful way. Knowing this is critical to understanding whether the results of the scaling laws study are valid, given the examples given throughout the paper of scaling laws study conclusions changing depending on the process details. Porian et al. (2024) and Alabdulmohsin et al. (2022) use confidence intervals and goodness of fit measures to validate their scaling laws. Ghorbani et al. (2021) and Bansal et al. (2022) also do this. Otherwise, a majority of the papers that we report as validating their scaling laws mainly extrapolate to models a few orders of magnitude larger and observe the adherence to the scaling law obtained.

3.1.9 Our Replications and Analyses

Each of the choices discussed above in Sections 3.1.3 - 3.1.8 may have a crucial impact on the result, yet it remains common to critically underspecify the setup for fitting a power law. Scaling law works often fail to open source their model and code, making reproduction infeasible, and likely contributing to contradictory conclusions as discussed in Section 3.1.2. Though some efforts have been made (Porian et al., 2024; Besiroglu et al., 2024) to reconcile such discrepancies, there is still only sparse understanding of the impact of each of the decisions we discuss. To investigate the significance of these scaling law optimization decisions, we vary these choices to fit our own scaling laws. We fit both the Chinchilla-scraped data from Besiroglu et al. (2024), and data from our own models.

Reconstructed Chinchilla data (Besiroglu et al., 2024) This data is extracted from a vector-based figure in the pdf of Hoffmann et al. (2022), who claim that this includes all models trained for the paper. It consists of 245 datapoints, each corresponding to a final checkpoint collected at the end of model training. There is a potential risk of errors in this recovery process.

Data from Porian et al. (2024) This data includes training losses for Transformer LMs ranging in size from 5 to 901 million parameters, each trained on OpenWebText2 (Gao et al., 2020) and RefinedWeb data (Penedo et al., 2023), across a variety of data and compute budgets, from 5 million to 14 billion tokens, depending on parameter count. Each model is trained with a different peak learning rate and batch size setting, found by fitting a separate set of scaling laws.

Our models We train a variety of Transformer LMs, ranging in size from 12 million to 1 billion parameters, on varied data and compute budgets and hyperparameter settings. Details about our setup, including hyperparameters, are listed in Appendix B.1. We open source all of our models, evaluation results, code, and FLOP calculator at `https://github.com/hadasah/scaling_laws`

We fit a multitude of power laws and study the effects of: (1) power law form; (2) model learning rate; (3) compute budget, model size, and data budget range and coverage; (4) definition of N and C; (5) inclusion of mid-training checkpoints; (6) power law parameter initialization; (7) choice of loss and (8) optimizer.

We plot all results in Appendix Figure B.1. In our initial analysis of (5), we find substantial differences, sometimes improvements, in stability of fit when including mid-training checkpoints. As a result, we include analyses and corresponding figures for power law fitting with only data from final checkpoints, as well as with all mid-training checkpoints, when applicable. Based on our observations, we also make some more concrete recommendations in Appendix §B.4, with the caveat that following the recommendations cannot guarantee a good scaling law fit.

3.1.10 Form (§3.1.3)

We consider the (1) baseline Hoffmann et al. (2022) form (Approach 3) and then (2) apply the trick employed by Muennighoff et al. (2024) of setting $\alpha = \beta$ in the Hoffmann et al. (2022) form, which assumes the optimal D/N ratio stays roughly constant – $L(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\alpha}$. We also compare with (3) an ISOFlop approach (Approach 3 of Hoffmann et al. (2022)), in which we fit an optimal N for each compute budget C, $N_{opt}(C)$, which can then be used to fit the predicted

loss $L(N, D)$. This approach usually necessitates the usage of mid-training checkpoints (discussed further in §3.1.12), as it is infeasible to train a large enough number of models for each FLOP budget considered. However, we apply it here without using only final model checkpoints, and extend to mid-training checkpoints in §3.1.12). We adapt the implementation from Porian et al. (2024), which contains more details about interpolation of data points and specific hyperparameters. In all data sets, (2) approaches the law reported by Hoffmann et al. (2022), but (1) only does so for the data from Porian et al. (2024) (Figure B.1a).

We experiment with the power law form reported by Kaplan et al. (2020), but this consistently yields a law which suggests that the optimal number of data points is 1, even when varying many aspects of the power law fitting procedure. The difficulty of fitting to this form might be partially a result of severe under-reporting in Kaplan et al. (2020) with regard to procedural details, including hyperparameters for both model training and fitting.

3.1.11 Training (§3.1.6)

Hu et al. (2024) study the effects of several model training decisions, including batch size, learning rate schedule, and model width. Their analysis focuses on optimizing hyperparameters, not on the ways hyperparameter and architecture choices affect the reliability of scaling law fitting. Observed variations between settings suggest that suboptimal performance could skew the scaling law fit.

To substantiate this further, we simulate the effects of not sweeping the learning rate in our models. As a baseline, (1) we sweep at each (N, D) pair for the optimal learning rate over a range of values, at most a multiple of 2 apart. Next, (2) we use a learning rate of 1e-3 for all N, the optimal for our 1 billion parameter models, and do the same for (3) 2e-3 and (4) 4e-3, which is optimal for our 12 million parameter models. Lastly, we use all models across all learning rates at the same N and D. Results vary dramatically across these settings. Somewhat surprisingly, using all learning rates results in a very similar power law to sweeping the learning rate, whereas using a fixed learning rate of 1e-3 or 4e-3 yields the lowest optimization loss or closest match to the Hoffmann et al. (2022) power laws, respectively (Figure B.1b).

We also study the effects of limiting model scale range, data scale range (implicitly), and

data-to-parameters range by filtering all datasets: we compare (1) using all N, D scales, (2) only models with N up to about 100 million or (3) 400 million parameters, and including (4) only models with $D/N \leq 18$ or (5) $D/N \geq 22$. These ranges are designed to exclude $D/N = 20$, the rule of thumb based on Hoffmann et al. (2022). The minimum or maximum D/N ratio tested does skew results; above 10^{22} FLOPs, (4) and (5) fit to optimal ratios $D/N < 18$ and $D/N > 22$, respectively. Removing our largest models in (2) also creates a major shift in the predicted optimal D/N (Figure B.1c).

Related to choice of N, D and C values, we investigate different ways of counting N and C, specifically whether to include embedding parameters and FLOPs. We compare (1) our baseline, which includes embeddings in both N and C, with (2) excluding embeddings only in N, (3) excluding embeddings only in C, (4) excluding embeddings in both N and C. We also compare to (5) using the $C = 6ND$ approximation, including embedding parameters. In all other settings, we calculate the FLOPs in a manner similar to Hoffmann et al. (2022), and we open source the code for these calculations. With both datasets, the exclusion of embeddings in FLOPs has very little impact on the final fit. Similarly, using the $C = 6ND$ approximation has no visible impact. For the Porian et al. (2024) models, the exclusion of embedding parameters in the calculation of N results in scaling laws which differ substantially, and with increasing divergences at large scales (Figure B.1d).

3.1.12 Data Collection (§3.1.7)

We compare using data from (1) only the final checkpoints of each model with (2) using all available mid-training checkpoints. We also consider removing checkpoints collected during the (3) first 10%, (4) 20%, and (5) 50% of training, Using mid-training checkpoints sometimes results in more stable fits which are similar to the Hoffmann et al. (2022) scaling laws, but the effect is unreliable and is dependent on other decisions. Following this finding, we also run all other analyses using setting (2), and find that these scaling law fits are often more consistent when varying other aspects of the fitting procedure (Figure B.1e).

3.1.13 Fitting (§3.1.8)

We vary the initialization method for our power law fitting procedure: (1) our baseline replication of Hoffmann et al. (2022) Approach 3, which conducts the full optimization process on a grid of 6x6x5x5x5=4500 initializations (Hoffmann et al., 2022), (2) searching for the lowest loss initialization point (Caballero et al., 2022) and optimizing only from that initialization, (3) optimizing from the 1000 lowest loss initializations, (4) randomly sampling k(=100) points (Frantar et al., 2023; Tao et al., 2024), and (5) initializing with the coefficients found in Hoffmann et al. (2022), as Besiroglu et al. (2024) does. With the Besiroglu et al. (2024) data, (5) yields a fit nearly identical to that reported by Hoffmann et al. (2022), although (1) results in the lowest fitting loss. With the Porian et al. (2024) data, all approaches except (2) yield a very similar fit, which gives a recommended D/N ratio similar to that of Hoffmann et al. (2022). However, using our data, (2) optimizing over only the most optimal initialization yields the best match to the Hoffmann et al. (2022) power laws, followed by (5) initialization from the reported Hoffmann et al. (2022) scaling law parameters. Optimizing over the full grid yields the power law that diverges most from the Hoffmann et al. (2022) law, suggesting the difficulty of optimizing over this space, and the presence of many local minima (Figure B.1f).

Then, we analyze the choice of loss objectives, including (1) the baseline log-Huber loss, (2) MSE, (3) MAE, and (4) the Huber loss. We find that there is somewhat less variance in the resulting power laws than we have seen in other power law fitting decisions, but the recommended optimal data budgets still span a wide range (Figure B.1g).

Finally, we consider the choice of optimizer. We first fit a power law to using the original (1) L-BFGS, with settings matching those described by Hoffmann et al. (2022). L-BFGS and BFGS implementations have an early stopping mechanism, which conditions on the stability of the solution between optimization steps. We experiment with setting this threshold for L-BFGS to a (2) higher value 1e-4 (stopping earlier). We find that using any higher or lower values results in the same solutions or instability for all three datasets. L-BFGS and BFGS also have the option to use the true gradient of the loss, instead of an estimate, which is the default. In this figure, we include

this setting, (3) using the true gradient for L-BFGS. We compare these L-BFGS settings to (4) BFGS. We test the same tolerance value and gradient settings, and find that none of these options change the outcome of BFGS in our analysis. Finally, we compare to (5) non-linear least squares and (6) pure grid search, using a grid 5 times more dense along each axis as that which we used for initialization with other optimizers. This density is chosen to approximately match the runtime of L-BFGS. Many of these optimizer settings converge to similar solutions, but this depends on the data, and the settings which diverge from this majority do not follow any immediately evident pattern (Figure B.1h).

3.1.14 Conclusion

We survey over 50 papers on scaling laws, and discuss differences in form, training setup, evaluation, and curve fitting, which may lead to significantly different conclusions. We also discuss significant under-reporting of details crucial to replicating the conclusion of these studies, and provide guidelines in the form of a checklist aid researchers in reporting more complete details. In addition to discussing several prior replication studies in literature, we empirically demonstrate the fragility of this process, by systematically varying these choices on available checkpoints and models that we train from scratch. We choose to avoid overly prescriptive recommendations, because there is no known set of actions which can guarantee a good scaling law fit, but we make some suggestions based on patterns in our findings (Appendix §B.4). Despite our preliminary investigations, our understanding of which decisions may skew the results of a scaling law study is sparse, and defines the path for future work.

Paper	Curve-fitting Method	Loss Objective	Hyperparameters Reported?	Initialization	Are scaling laws validated?
Rosenfeld et al. (2019)	Least Squares Regression	Custom error term	N/A	Random	Y
Mikami et al.	Non-linear Least Squares in log-log space		N/A	N/A	Y
Schaeffer et al. (2023)	NA	NA	NA	NA	NA
Sardana & Frankle (2023)	L-BFGS	Huber Loss	Y	Grid Search	N
Sorscher et al. (2022)	NA	NA	NA	NA	NA
Caballero et al. (2022)	Least Squares Regression	MSLE	N/A	Grid Search, optimize one	Y
Besiroglu et al. (2024)	L-BFGS	Huber Loss	Y	Grid Search	Y
Gordon et al. (2021)	Least Squares Regression		N/A	N.S.	N
Bansal et al. (2022)	NS	NS	N	NS	N
Hestness et al. (2017)	NS	RMSE	N	NS	Y
Bi et al. (2024)	NS	NS	N	NS	Y
Bahri et al. (2024)	NS	NS	N	NS	N
Geiping et al. (2022)	Non-linear Least Squares		NA	Non-augmented parameters	Y
Poli et al. (2024)	NS	NS	N	NS	N
Hu et al. (2024)	scipy curvefit	NS	N	NS	N
Hashimoto (2021)	Adagrad	Custom Loss	Y	Xavier	Y
Ruan et al. (2024)	Linear Least Squares	Various	N/A	N/A	Y
Anil et al. (2023)	Polynomial Regression (Quadratic)	N.S.	N	N.S.	Y
Pearce & Song (2024a)	Polynomial Least Squares	MSE on Log-loss	N/A	N/A	N
Cherti et al. (2023)	Linear Least Squares	MSE	N/A	N/A	N
Porian et al. (2024)	Weighted Linear Regression	weighted SE on Log-loss	N/A	N/A	Y
Alabdulmohsin et al. (2022)	Least Squares Regression	MSE	Y	N.S.	Y
Gao et al. (2024)	N.S	N.S	N.S	N.S	N.S
Muennighoff et al. (2024)	L-BFGS	Huber on Log-loss	Y	Grid Search, optimize all	Y
Rae et al. (2021)	None	None	N/A	N/A	N
Shin et al. (2023)	NA	NA	NA	NA	NA
Hernandez et al. (2022)	NS	NS	NS	NS	NS
Filipovich et al. (2022)	NS	NS	NS	NS	NS
Neumann & Gros (2022)	NS	NS	NS	NS	NS
Droppo & Elibol (2021)	NS	NS	NS	NS	NS
Henighan et al. (2020)	NS	NS	NS	NS	NS
Goyal et al. (2024)	Grid Search	L2 error	Y	NA	Y
Aghajanyan et al. (2023)	L-BFGS	Huber on Log-loss	Y	Grid Search, optimize all	Y
Kaplan et al. (2020)	NS	NS	NS	NS	N
Ghorbani et al. (2021)	Trust Region Reflective algorithm, Least Squares	Soft-L1 Loss	Y	Fixed	Y
Gao et al. (2023)	NS	NS	NS	NS	Y
Hilton et al. (2023)	CMA-ES+Linear Regression	L2 log loss	Y	Fixed	Y
Frantar et al. (2023)	BFGS	Huber on Log-loss	Y	N Random Trials	Y
Prato et al. (2021)	NS	NS	NS	NS	NS
Covert et al. (2024)	Adam	Custom Loss	Y	NS	Y
Hernandez et al. (2021)	NS	NS	NS	NS	Y
Ivgi et al. (2022)	Linear Least Squares in Log-Log space	MSE	NA	NS	Y
Tay et al. (2022a)	NA	NA	NA	NA	NA
Tao et al. (2024)	L-BFGS, Least Squares	Huber on Log-loss	Y	N Random Trials from Grid	Y
Jones (2021)	L-BFGS	NS	NS	NS	NS
Zhai et al. (2022)	NS	NS	NS	NS	NS
Dettmers & Zettlemoyer (2023)	NA	NA	NA	NA	NA
Dubey et al. (2024)	NS	NS	NS	NS	Y
Hoffmann et al. (2022)	L-BFGS	Huber on Log-loss	Y	Grid Search, optimize all	Y
Ardalani et al. (2022)	NS	NS	NS	NS	NS
Clark et al. (2022)	L-BFGS-B	L2 Loss	Y	Fixed	NS

Table 3.2: We provide an overview of which papers provide specific details required to reproduce how they fit their scaling law equation.

Chapter 4

ATLAS: ADAPTIVE TRANSFER SCALING LAWS FOR MULTILINGUAL PRETRAINING, FINETUNING, AND DECODING THE CURSE OF MULTILINGUALITY

4.1 Introduction

Scaling laws research has focused overwhelmingly on English—yet the most prominent AI models explicitly serve billions of international users. In this work, we undertake the largest multilingual scaling laws study to date, totaling 774 multilingual training experiments, spanning 10M-8B model parameters, 400+ training languages and 48 evaluation languages. We introduce the ADAPTIVE TRANSFER SCALING LAW (ATLAS) for both monolingual and multilingual pretraining, which outperforms existing scaling laws' out-of-sample generalization often by more than 0.3 R^2. Our analyses of the experiments shed light on multilingual learning dynamics, transfer properties between languages, and the curse of multilinguality. First, we derive a cross-lingual transfer matrix, empirically measuring mutual benefit scores between $38 \times 38 = 1444$ language pairs. Second, we derive a language-agnostic scaling law that reveals how to optimally scale model size and data when adding languages without sacrificing performance. Third, we identify the computational crossover points for when to pretrain from scratch versus finetune from multilingual checkpoints. We hope these findings provide the scientific foundation for democratizing scaling laws across languages, and enable practitioners to efficiently scale models—beyond English-first AI.

Scaling laws research has focused overwhelmingly on English (Kaplan et al., 2020; Hoffmann et al., 2022; Li et al., 2025a). However, most of the major closed and open models now explicitly target massively multilingual uses (OpenAI, 2025; Anthropic, 2025; Google DeepMind, 2025; DeepSeek-AI, 2024; Team OLMo et al., 2024). Nonetheless, multilingual scaling laws research in the public sphere is limited. Among proprietary lab reports, only Llama-3 briefly discuss their multilingual scaling laws, though they train on only 8% non-English tokens (Dubey et al., 2024).

Prominent public work investigates scaling laws for data mixing (Goyal et al., 2024; Ye et al., 2024; Ge et al., 2024a), scaling laws for machine translation (Fernandes et al., 2023; Gordon et al., 2021), multilingual instruction tuning/adaptation (Shaham et al., 2024; Lai et al., 2024; Weber et al., 2024), and only a couple of recent rigorous contributions investigate scaling for smaller multilingual models (<$45M$ and <$1.2B$ respectively) (Chang et al., 2024; He et al., 2024).

Extending these prior works, we seek to fill the ample remaining knowledge gaps with comprehensive examinations of the following research questions. First, we explore how the properties of different languages' scaling laws differ. Second, we measure the cross-lingual transfer benefits between 38 languages. To our knowledge, this represents the most comprehensive available empirical resource to explicitly measure language transfer across $38 \times 38 = 1444$ language pairs. Third, we succeed in modeling the *curse of multilinguality*—the phenomena where adding languages to the training mixture can degrade loss for each language, due to limited model capacity. Lastly, we measure when it is more efficient to pretrain from scratch, or finetune from a general-purpose multilingual checkpoint, resolving an unaddressed practical question. In pursuing these research questions, we develop new tools and estimates for multilingual practitioners, as well as the ADAPTIVE TRANSFER SCALING LAW, which significantly outperforms prior work across several dimensions of generalization. In total, our pretraining and finetuning experiments span 774 independent experiments on MADLAD-400 (Kudugunta et al., 2024), for models sized $10M - 2B$, and evaluating 48 languages. Our main contributions include:

1. **The ADAPTIVE TRANSFER SCALING LAW** that offers better multilingual generalization to larger/unseen N, D, C, and training mixes M, than prior work (Table 4.1). In Figure 4.1 we also estimate a compute efficiency tax between multilingual and monolingual training.

2. **A** 38×38 **CROSS-LINGUAL TRANSFER MATRIX**, providing the largest resource for empirically measured transfer benefits/interference between languages (Figure 4.2).

3. **A scaling law for the *curse of multilinguality***, that informs practitioners how much to scale (N, D) to accommodate expansions in a models' language coverage (Figure 4.5).

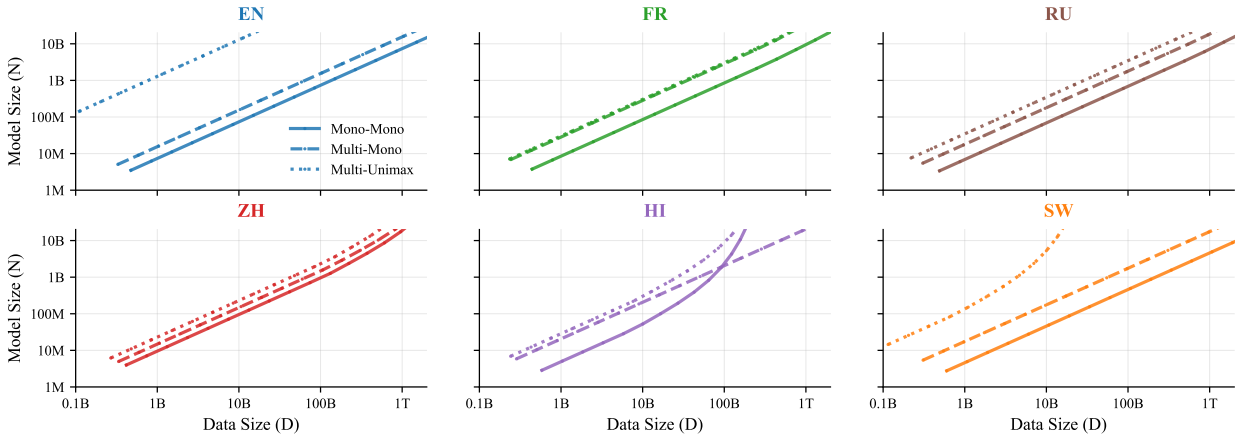


Figure 4.1: **Optimal Scaling Trajectories** for English, French, Russian, Chinese, Hindi, and Swahili. The law for [monolingual vocabulary, monolingual training]=(—), the law for [multilingual vocabulary, monolingual training]=(- - -), and the law for [multilingual vocabulary, unimax training]=(···). **We find (1) per-language optimal scaling trajectories are similar, (2) there is a compute efficiency tax for training with *multilingual* vocabularies or training sets (especially for English), and (3) as Hindi and Swahili observe data repetition their curves slope upward from diminishing returns.**

4. **A general pretrain vs finetune formula**, that informs practitioners whether it is more efficient to pretrain from scratch or begin from a multilingual Unimax checkpoint (Figure 4.7).

4.1.1 Experimental Setup

Datasets and Evaluation We use the MADLAD-400 dataset (Kudugunta et al., 2024), a popular CommonCrawl-based dataset with the most expansive coverage of languages, totaling over 400. The MADLAD-400 authors prioritized multilingual-specific curation for this pretraining corpus, including language detection, filtering, and preprocessing. For 50 languages, chosen to represent a range of language families, scripts, and degrees of resourcefulness, we partition a random test set, to evaluate vocabulary-insensitive loss (Tao et al., 2024) for fairer cross-lingual comparisons. Following (Muennighoff et al., 2024; He et al., 2024), we evaluate the fit of our parametric scaling laws using R^2 calculated on held-out test sets partitioned to assess specific dimensions of

generalization: $R^2(N)$ for the largest model sizes, $R^2(D)$ for the largest token ranges, $R^2(C)$ for the largest compute runs, and $R^2(M)$ for unique training language mixtures not seen at training (more details in Subsection C.2.3). We separate these dimensions in response to prior work that has demonstrated the importance of rigorous test-set splits in scaling laws research (Li et al., 2025a).

Model Training We pretrain models from 10M to 8B parameters, using hyperparameter choices similar to those trained in Kudugunta et al. (2024) (with updates as prescribed by recent work). In this work, we train both monolingual and multilingual models, varying the model size N, training tokens D, multilingual data proportions, and the total training compute C. In particular, we focus on training monolingual models, bilingual models, and massively multilingual Unimax models (Chung et al., 2023). most experiments center around scale, token, and mixture variations on these languages: English, French, Chinese, Hindi, Swahili, Russian, Spanish, Portuguese, Japanese, though certain experiments evaluate across 50 languages. We use a 64k Sentence Piece Model vocabulary (Kudo, 2018). Across experiments we pretrain 280 monolingual models, 240 bilingual models, 120 multilingual mixtures, and we finetune 130 monolingual models from Unimax checkpoints, totaling over 750 independent training runs. Full experimental details, including the specific language choices, hyperparameters, and training mixtures are provided in full in Section C.2.

4.1.2 Adaptive Scaling Laws for Monolingual & Multilingual Settings

Challenges with existing scaling laws for multilingual modeling. In this section, we discuss our attempts to precisely model different monolingual and multilingual pretraining experiments. Scaling laws for English are typically based on Chinchilla (Hoffmann et al., 2022)—which we refer to as the CHINCHILLA SCALING LAW (CSL). Hoffmann et al. (2022)'s two power-law terms (Equation (4.1)), allows us to separate how loss scales model size N and data size D. E is the irreducible loss (representing the entropy of natural text), (A, B) are learned coefficients and and (α, β) are scaling exponents for for model size and data size respectively. Beyond English, languages may have different scaling laws (Arnett & Bergen, 2025), and language resources are often limited, requiring a DATA-CONSTRAINED SCALING LAW (DCSL) (Muennighoff et al., 2024) to account for diminishing returns after multiple epochs. However, DCSL requires ample data both before and

Table 4.1: The R^2 evaluation metrics for the fitted scaling laws, holding out separate dimensions of generalization: the largest model sizes N, most training tokens D, most compute C, and unique multilingual training mixtures M. In both monolingual and multilingual settings, we average R^2 across languages=[EN, FR, RU, ZH, HI, SW], including [ES, DE] for the multilingual setting. **We find ADAPTIVE TRANSFER SCALING LAW outperforms prior work in Monolingual and Multilingual settings.** We ablate the use of the terms for D_{other} and transfer languages ($\sum_{\mathcal{K}_t} D_i$).

	SCALING LAWS	R^2	$R^2(N)$	$R^2(D)$	$R^2(C)$	$R^2(M)$
MONO.	CHINCHILLA SCALING LAW (Hoffmann et al., 2022)	0.94	0.68	0.94	0.90	–
	DATA-CONSTRAINED SCALING LAW (Muennighoff et al., 2024)	0.93	0.78	0.93	0.88	–
	[Ours] ATLAS (D_t ONLY)	0.92	**0.88**	0.91	0.88	–
MULTI.	CHINCHILLA SCALING LAW (Hoffmann et al., 2022)	0.64	-0.99	0.72	0.66	0.61
	MULTILINGUAL SCALING LAW (He et al., 2024)	0.67	-0.65	0.73	0.67	0.70
	[Ours] ATLAS (D_t ONLY)	0.70	-0.75	0.80	0.72	0.64
	[Ours] ATLAS ($D_t + D_{\text{other}}$)	**0.98**	**0.89**	**0.97**	**0.97**	0.66
	[Ours] ATLAS ($D_t + D_{\text{other}} + \sum_{\mathcal{K}_t} D_i$)	**0.98**	**0.89**	**0.96**	**0.98**	**0.82**

after one epoch to accommodate its two-stage fitting process. For high-resource languages (English, French, Chinese), it is costly to collect ample data beyond one epoch, and for low/mid-resource languages (even Hindi, Hebrew, and Swahili) it can be very difficult to collect sufficient observations below one epoch.[1] As such, even for monolingual modeling, we require a scaling law that is data repetition-aware, and robust to different collection allowances for (N, D).

For modeling multilingual mixtures, monolingual scaling laws can be used (either with D representing the total or target language data), or He et al. (2024) introduce MULTILINGUAL SCALING LAW (MSL), which expresses the loss for target language t using N, D and the sampling ratio for the target languages' language family in the training mixture. However, each of these

[1] In MADLAD-400 these languages have $< 0.3\%$ English tokens. At standard batch sizes, even frequently evaluating (every $1k$) isn't enough to collect many (sometimes any) observations before 1 epoch is reached.

solutions only accounts for one of: the sampled target language data D_t, the total data altogether $D_t + D_{\text{other}}$, or a set of the target and likely positive transfer languages $D_t + \sum_i D_i$. We posit that a better model would separate and learn to weight these independent contributions. In summary, none of the existing multilingual solutions account for (a) multi-epoch data repetition, or (b) the cross-lingual transfer effects beyond the target's language family.

$$\mathcal{L}(N, \mathcal{D}_{\text{eff}}) = E \; + \; \frac{A}{N^{\alpha}} \; + \; \frac{B}{\mathcal{D}_{\text{eff}}^{\beta}} \tag{4.1}$$

$$\mathcal{D}_{\text{eff}} = \underbrace{\mathcal{S}_{\lambda_t}(D_t; U_t)}_{\text{Monolingual}} + \underbrace{\sum_{i \in \mathcal{K}} \tau_i \, \mathcal{S}_{\lambda_i}(D_i; U_i)}_{\text{Transfer Languages}} + \underbrace{\tau_{\text{other}} \, \mathcal{S}_{\lambda_{\text{other}}}(D_{\text{other}}; U_{\text{other}})}_{\text{Other Languages}} \tag{4.2}$$

$$\mathcal{S}_{\lambda}(D; U) = \begin{cases} D, & D \leq U \quad (\leq 1 \text{ epoch}) \\ U \left[1 + \frac{1 - \exp(-\lambda(D/U - 1))}{\lambda} \right], & D > U \quad (> 1 \text{ epoch}) \end{cases} \tag{4.3}$$

The ADAPTIVE TRANSFER SCALING LAW. To resolve these challenges, we introduce the ADAPTIVE TRANSFER SCALING LAW (ATLAS), a simpler variant of DCSL that is repetition-aware, introduces fewer additional parameters (for the monolingual variant), is fit in one stage, and is adaptable to an open-ended number of languages that would benefit from an explicit cross-lingual term. In Equation (4.1) the core scaling law formula includes the standard parameters governing the irreducible loss and N, D scaling: E, A, B, α, β. Its effective data exposure term D_{eff} (Equation (4.2)) unpacks data sources into three terms: a Monolingual term for the target language data D_t, an optional Transfer Language term for up to $|\mathcal{K}_t|$ languages we wish to learn independent transfer coefficients for τ_i, and an Other Languages remainder term that sums all training tokens not already accounted for in the first two terms $D_{\text{other}} = D_{\text{tot}} - D_t - \sum_{\mathcal{K}_t} D_i$. The latter two terms are optionally added for multilingual modeling, and $\mathcal{K}_t$ can be adapted as any combination of the languages with presumed highest positive transfer to the target language, or languages with the greatest representation in the experiment training mixture. We found the latter was most effective and selected the $|\mathcal{K}_t| = 3$ most highly co-sampled languages along target language t across mixtures. For each term, we apply a saturation function (Equation (4.3)), ensuring a smooth decay on the

effective data for each subsequent epoch where the training tokens have surpassed that languages' unique tokens U. The new parameters introduced over standard scaling laws are: the repetition parameter λ, which is shared across each data source, the transfer weights τ_i for each language in $\mathcal{K}_t$ (which are initialized from language transfer scores derived in Subsection 4.1.3), and the transfer weight τ_{other} for the remainder of language tokens.

Research Question: *How do scaling laws differ by language, and by monolingual vs multilingual training mixtures?*

Monolingual scaling behavior is consistent in form, and multilingual variants exhibit a variable-sized compute-efficiency tax. Figure 4.1 compares optimal scaling trajectories across six-variable resourced languages, under three regimes: *monolingual vocabulary + monolingual training*, *multilingual vocabulary + monolingual training*, and *multilingual vocabulary + unimax training*. Across languages, the monolingual curves are nearly parallel, indicating similar exponents and comparable returns to additional data and parameters (see full law parameters in Table C.6). Both multilingual vocabulary and Unimax training shift the frontier upwards, evidencing a *compute-efficiency tax* relative to the monolingual-vocabulary/monolingual-training setting. This is most pronounced for English, indicating it benefits less from language transfer than other languages do from English. For Hindi and Swahili, the right tail bends upward, consistent with diminishing returns from severe data repetition after many epochs. These qualitative trends motivate a single functional form that remains stable across languages while accounting for vocabulary/training-regime effects.

Research Question: *How well can our scaling laws capture unique monolingual constraints, and complex multilingual cross-lingual transfer dynamics?*

ATLAS outperforms prior scaling laws, with a more robust fit across held-out axes in monolingual and multilingual settings. Table 4.1 evaluates fitted laws by holding out the largest model sizes N, token ranges D, compute $C = 6ND$, and held-out language mixtures M, reporting R^2 averaged over available languages. In the monolingual setting, all scaling laws perform comparably across most dimensions of generalization. However, when generalizing to the largest size of model, laws that model data repetition outperform those that don't. ATLAS provides the strongest generalization to greater model sizes: $R^2(N) = \mathbf{0.88}$ versus 0.68 (CSL) and 0.78 (DCSL), respectively.

In the multilingual setting, monolingual scaling laws can only observe their target data tokens, and as such perform adequately, but not well. In particular, they are unable to generalize to the largest models $R^2(N)$ at all. This is also the case for MSL, although it obtains better generalization to unseen language mixture in $R^2(M) = 0.69 > 0.64$ by virtue of modeling the whole language family. By separating the sources of data, ATLAS is able to significantly outperform the other laws across all dimensions, frequently achieving > 0.9 fit. However, our ablation of the data terms shows that only by adding the Transfer Language term is the model able to outperform MSL and achieve a better multilingual mixture generalization $R^2(M) = 0.82$. In summary, ATLAS offers a simple framework to adaptively model cross-lingual transfer, and enables reliable extrapolation across out-of-sample dimensions in both monolingual and multilingual settings—addressing a critical gap where existing approaches fall short.

4.1.3 How Do Languages Benefit or Interfere with Each Other?

Research Question: *Which languages synergize or interfere most with one another's performance?*

When allocating multilingual training mixtures, practitioners need to know which *source* language(s) provide the most positive benefit, in co-training, to the *target* language(s) they are optimizing for. Throughout this paper, we use the terms transfer or interference to refer to the positive (or negative) inductive transfer between the training signal of two or more tasks (Caruana, 1997). We construct, to our knowledge, the most comprehensive matrix of language-to-language transfer scores, to assist practitioners in selecting training languages.

$$\text{Language Transfer Score } (s \to t) = \text{BTS}_{s \to t} = -\frac{\sigma_{\text{bi}}(L_{\text{t}}(d_{\text{mono}})) - 2d_{\text{mono}}}{d_{\text{mono}}}$$

To do this, we measure how much training in a source language s affects the test loss on a target language t. We define a Bilingual Transfer Score (BTS) as the relative training efficiency of a bilingual model $(s, t) = (50\%, 50\%)$ compared to the monolingual model t at reaching the same loss level.[2] d_{mono} denotes a pre-defined target step (42B tokens), $L_t(d)$ computes the monolingual

[2]We measure $\text{BTS}_{s \to t}$ directly for 80 language pairs, then estimate it using other training signals (with high fidelity $R^2 = 0.85$), as detailed in Subsection C.2.6.

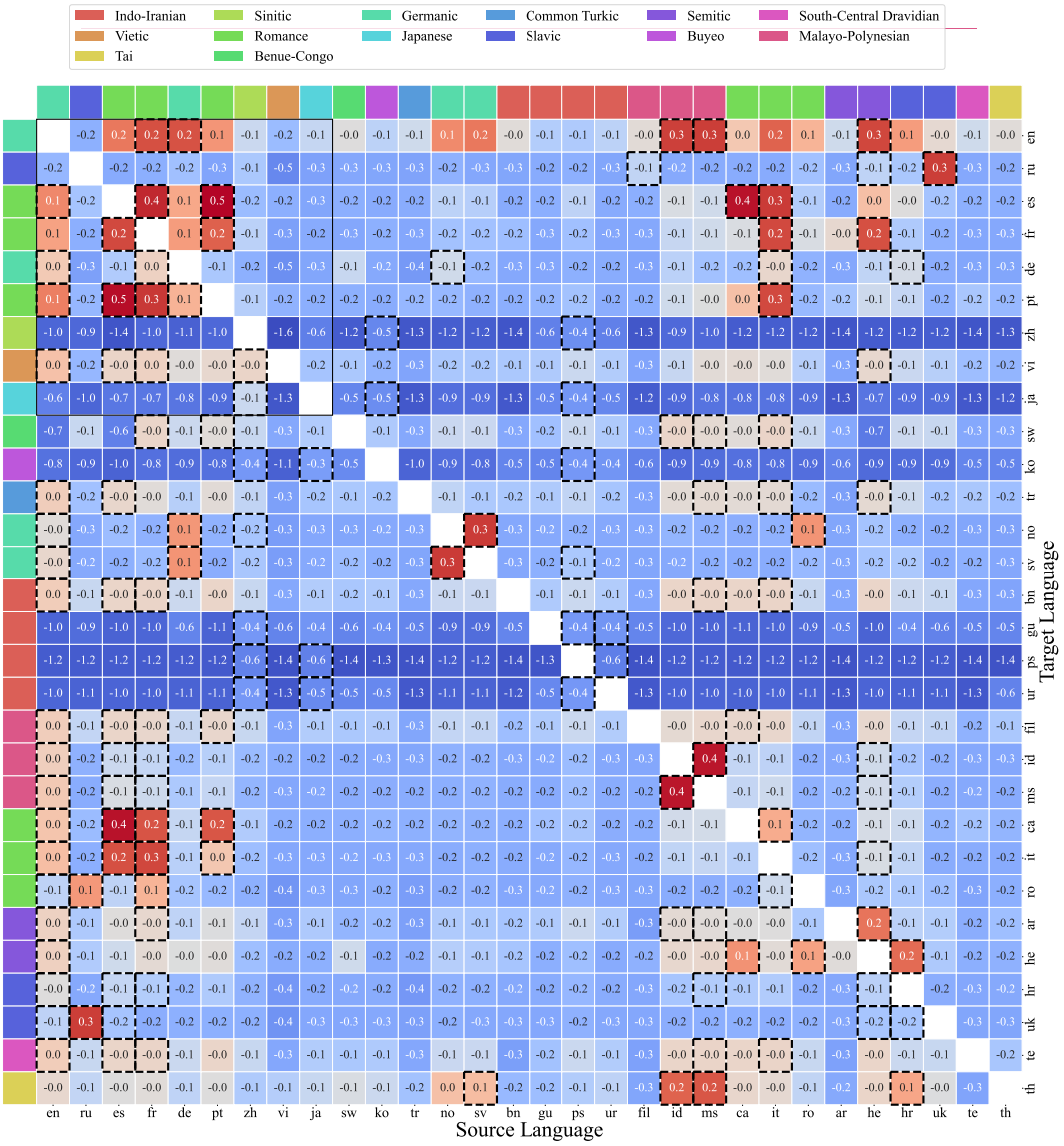


Figure 4.2: The CROSS-LINGUAL TRANSFER MATRIX, depicting the measured Language Transfer Score across 30 × 30 language pairs. Positive scores indicate more positive transfer, negative scores more interference, during bilingual co-training. The dashed boxes indicate the top-5 source languages for each target language. Refer to Appendix C.2.6 for full details, and Figure C.2 for the larger 38 × 38 matrix. **While English is the best source language for many of the languages, we find language similarity is highly predictive of these scores.**

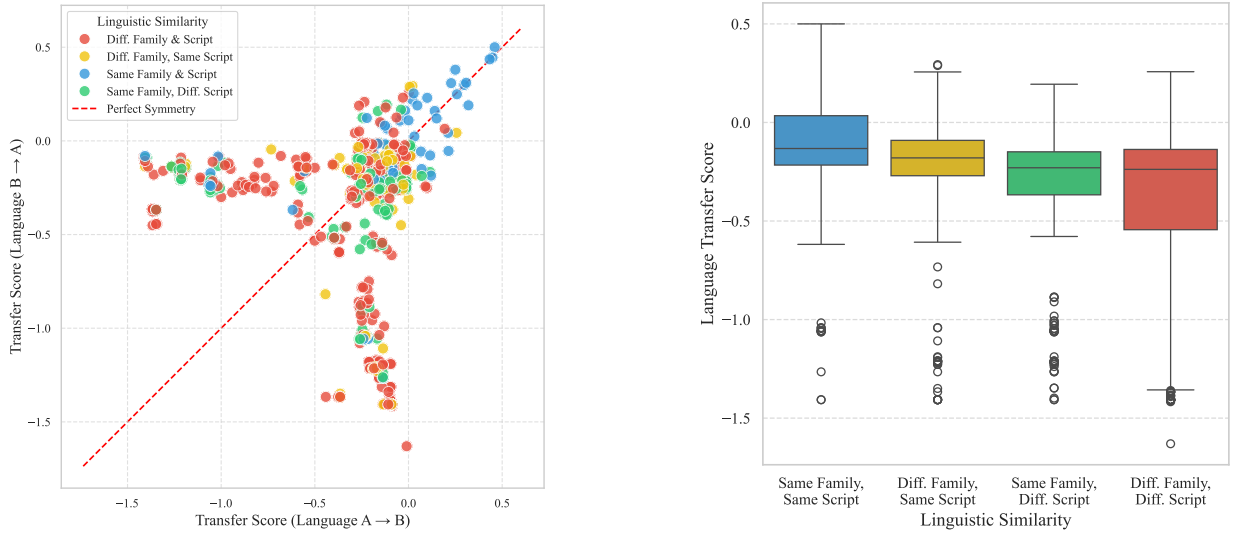


Figure 4.3: **Left:** A language transfer scatter plot, comparing score symmetry: is language A as helpful to B as vice versa? Points cluster near the diagonal, indicating strong symmetry. **The most synergistic and symmetric pairs almost exclusively share both language family and script.** Greater linguistic distance correlates with increased asymmetry and reduced positive transfer. **Right:** The impact of linguistic similarity (via language family or script) on transfer scores. The box spans the inter-quartile range, with the median line at the center. **We find the differences between each group are statistically significant** ($p < .001$)**, suggesting that sharing either a language family or a script independently contribute greater positive cross-lingual transfer.**

models' loss at step d, and σ_{bi} finds the number of tokens for the bilingual model to reach that loss. Because $\text{BTS}_{s \to t} > 0$ is the zero-centered, when it is $= 0$ there is no transfer, > 0 there is positive transfer, and < 0 there is negative interference, leading to the bilingual model taking more than double the steps d_{mono} the monolingual model took to reach the target loss $L_{\text{t}}(d_{\text{mono}})$. Refer to Subsection C.2.5 for full experimental details and methodology.

In Figure 4.2, we plot the normalized BTS scores between 30×30 language pairs (or the full 38×38 in Figure C.2), spanning language families and scripts, on 2B parameter models. Where prior work has developed comprehensive resources for measuring syntactic/phonological distances between languages (Khan et al., 2025), or measured transfer scores with smaller models specifically between high and low-resource languages (Protasov et al., 2024), our work offers among the most expansive and rigorous, fully-symmetric transfer matrices, to our knowledge. It contains many compelling observations. For instance, as one might expect, notable positive transfer exists between related languages, such as between Spanish, Catalan, and Portuguese, or Swedish and Norwegian, or Indonesian and Malay. The top-5 most helpful source languages to co-train with per target are highlighted, with English appearing the most (19 of 30 instances), followed by French (16 of 30), Spanish (13 of 30), and even Hebrew (11 of 30). In other cases, low-resource languages such as Urdu or Pashto have significant negative transfer with all other languages considered.

Language script, followed by language family are strongly correlated with positive transfer scores. In Figure 4.3, we correlate the transfer scores from the full language matrix Figure 4.2. Our analysis confirms a clear and statistically significant link between transfer performance and the similarity of the source and target language. Specifically, we found sharing a script or family introduced statistically significant shifts in the mean of the transfer score (always $p < .001$). These findings corroborate (He et al., 2024)'s scaling laws that grouped data by language family. This effect is strongest for shared scripts. As shown in Figure 4.3, language pairs that share the same writing system (e.g. Latin) exhibit dramatically improved transfer scores compared to pairs with disparate scripts (e.g., Latin and Cyrillic), with a mean score of -0.23 versus -0.39. The larger effect size for script suggests that the ability to share surface-level representations and subword vocabularies is a primary mechanism for positive transfer, more so than deeper grammatical or

lexical similarities alone. In Subsection 4.1.4 and Subsection C.3.2 we explore how N, D affect language transfer.

Language transfer scores are often symmetric within the same language family and script, but surprisingly, they cannot be assumed to be reciprocal otherwise. Next, we examine the extent to which language transfer is symmetric, that is, whether the benefit from language A to B is correlated to that from B to A. As illustrated in Figure 4.3 (left), we find a surprising triangle pattern, with a Pearson correlation of $r = -0.11$. This implies two things. First, and perhaps most importantly, across all language pairs, there is no significant correlation, and because we observe language A is helpful to language B, *we cannot assume the same in reverse.* This corroborates findings in Li et al. (2025c) that altruistic languages don't always yield mutual benefits. Second, there does appear to be a clear structure to the scatter: language pairs that share family and script (blue) cluster mostly in the top right quadrant and more tightly around the identity line, whereas language pairs with differing script and family (red) are simultaneously less symmetric, and less synergistic. For instance, pairs (French, Spanish) and (Russian, Ukrainian) share highly symmetric transfer scores, whereas (Chinese, Farsi) and (Russian, Vietnamese) are highly asymmetric. We hope these findings, as well as the Figure 4.2 transfer matrix, will be directly applicable to practitioners selecting language mixtures based on empirical measurements, rather than intuition.

4.1.4 The Curse of Multilinguality—Model Capacity Constraints

Research Question: *Is the "curse of multilinguality" measurable—the phenomenon where adding languages to the training mixture can degrade loss of each language, due to limited model capacity?*

A model's capacity can hinder its ability to learn multiple languages at once—known as the *curse of multilinguality* (Conneau et al., 2019; Chang et al., 2024). We can empirically measure this relationship, between N, D, and the number of training languages K, by running experiments with variable K. Full experimental details and scaling law derivations are provided in Subsection C.2.7.

Modeling the Relationship between K, N, and D. In Figure 4.4 we examine the relative loss increase, over a monolingual baseline, from varying the number of training languages K, and either the model size N (left plot) or total training tokens D_{tot} (right plot). For simplicity, we don't

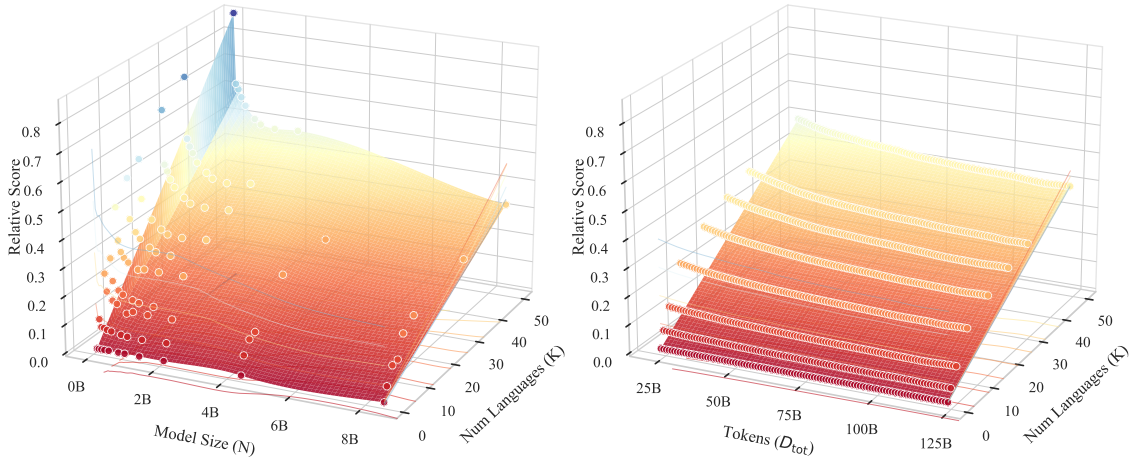


Figure 4.4: We empirically measure the relative degradation in loss (y-axis), as compared to a monolingual model of the same (N, D), from adding pretraining languages (z-axis). **Left:** We fix $D = 25B$ tokens, and vary the model size N. **Right:** We fix $N = 2B$ parameters, and vary the tokens D. The points are real empirical observations, averaged across languages. The mesh-grid is a surface estimate, using a cubic spline. **Target language loss is most affected by the number of training languages, but this loss penalty declines for larger models with more capacity.**

explicitly model token repetition, and we sample tokens from each language uniformly—we believe this is a reasonable assumption for models designed to serve all K languages. We observe three phenomena. First, the number of training languages K has far more impact on the relative loss than N or D. Relative loss appears to grow monotonically as K increases. Second, both increasing model size N and total tokens D_{tot} mitigate this penalty. However, computing the partial along the fitted surface shows $|\partial S/\partial \log N| > |\partial S/\partial \log D|$; indicating N delivers larger marginal gains than D_{tot}. Consequently, maintaining performance as K grows requires scaling both data and model size, but scaling the model size has a greater effect. To understand the relationship among these variables more precisely, we model per–target–language loss as

$$L(K, N, D_t) \;=\; L_\infty \;+\; A\,\frac{K^\phi}{N^\alpha} \;+\; B\,\frac{K^\psi}{D_t^\beta},$$

where, under even sampling the total tokens across languages are $D_{\text{tot}} = K\,D_t$. We tested several scaling law variations for this research question, but settled on this one as (a) it retains Chinchilla-style power-law decay properties that cleanly separate model capacity from data (it also reduces to Chinchilla when K=1), (b) it makes the role of the number of languages explicit and interpretable, and (c) it achieved a robust $R^2 \geq 0.87$, indicating a strong fit. The exponent ϕ captures how capacity requirements grow with the number of languages, while ψ captures how data requirements change with multilinguality. $\psi < 0$ indicates *positive transfer* (sublinear data needs per language as K increases), and $\psi > 0$ implies negative transfer/interference.

We fit the scaling law for each of 8 different languages, as well as the combination of all, achieving similar coefficients, and > 0.8 R^2 for each. Using the scaling law fit on all language data we find $\phi = 0.11$ and $\psi = -0.04$, indicating a mild capacity-driven curse of multilinguality, tempered by positive transfer across languages. In other words, as languages are added to the training mixture, a positive ϕ means the loss will decline from limited model capacity, however, a negative ψ means that less data per language is needed. (Though the total amount of data increases, as we sample from new languages too.) This provides clear, measurable evidence of positive cross-lingual transfer.

Estimating the Iso-Loss Frontier, and Compute-Optimal scaling of N, D when adding lan-

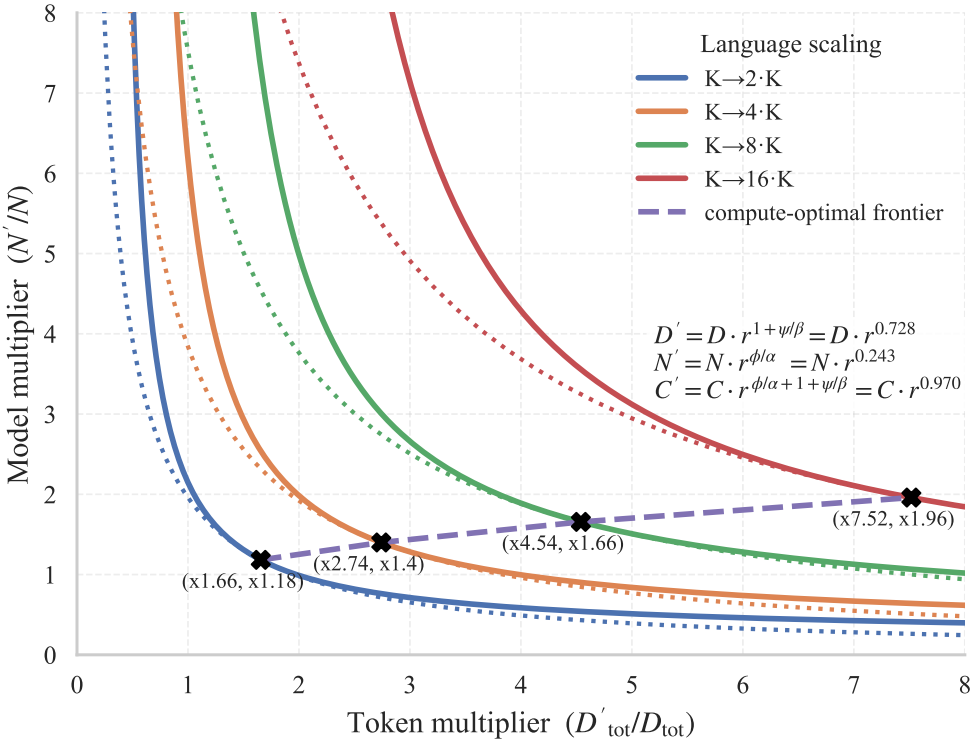


Figure 4.5: **Iso-loss from expanding language coverage:** When scaling the number of languages a model serves from $K \to r \cdot K$, we can estimate how much they need to increase model size N'/N and/or training tokens D'_{tot}/D_{tot}, without degrading any of their languages' loss. We derive the compute optimal scaling equations. **The *iso-losses* indicate that a practitioner should expand their compute budget by $C \cdot r^{0.97}$ to expand their language coverage by r.**

guages: $K \to rK$. Next, we examine the practical case where a practitioner wants to retrain a new model, increasing the number of languages it serves from K to rK, without hindering the performance the original model achieved on the existing languages. To accommodate these new languages, they will need to scale C with some combination of $N \to N'$ and $D_{\text{tot}} \to D'_{\text{tot}}$. We can compute this iso-loss by equating the loss terms: $L(K, N, D_{\text{tot}}) = \frac{AK^\phi}{N^\alpha} + \frac{BK^\psi}{D^\beta} = \frac{A(Kr)^\phi}{N'^\alpha} + \frac{B(Kr)^\psi}{D'^\beta}$. From this we can derive the closed-form expression for how to scale (N, D_{tot}) when increasing K to rK (as derived in Subsection C.2.7):

$$\frac{N^*(rK)}{N^*(K)} = r^{\phi/\alpha}, \qquad \frac{D_t^*(rK)}{D_t^*(K)} = r^{\psi/\beta}, \qquad \frac{D_{\text{tot}}^*(rK)}{D_{\text{tot}}^*(K)} = r^{1+\psi/\beta}.$$

We can also derive the minimal increase in the compute budget:

$$\frac{C'}{C} = \frac{N' D'_{tot}}{N D_{tot}} = \left(\frac{N'}{N}\right)^\star \left(\frac{D'_{\text{tot}}}{D_{\text{tot}}}\right)^\star = r^{1+\phi/\alpha+\psi/\beta}.$$

In Figure 4.5 we plot the iso-loss frontiers, and compute-optimal allocations, when scaling a model from K to $r \cdot K$ languages. It shows in closed-form how a practitioner should scale D_{tot}, N, and by extension their compute budget C to accommodate additional languages, without degrading their loss. For instance, we find expanding to $4 \cdot K$ languages requires a practitioner to expand D_{tot} by 2.74, and the model size N by 1.4. Incidentally, evenly sampling across languages, this actually corresponds to using sampling $1 - (2.74/4) = 32\%$ less data per language. Although there are fewer tokens in each target language (by 32%), the total tokens has risen by 2.74, and the positive transfer prevents any potential loss degradation.

4.1.5 Pretrain or Finetune?

Research Question: *When training a model for target language t, is it more effective to pretrain from scratch, or finetune a general-purpose massively multilingual checkpoint?*

Prior scaling law work focuses on compute efficiency with respect to pretraining from scratch. However, in reality, practitioners who aim to optimize for a target language t may have the option to start from multilingual public checkpoints or pretrain one multilingual checkpoint to serve as the starting point for many downstream models. Consider a practitioner with a compute budget C. In

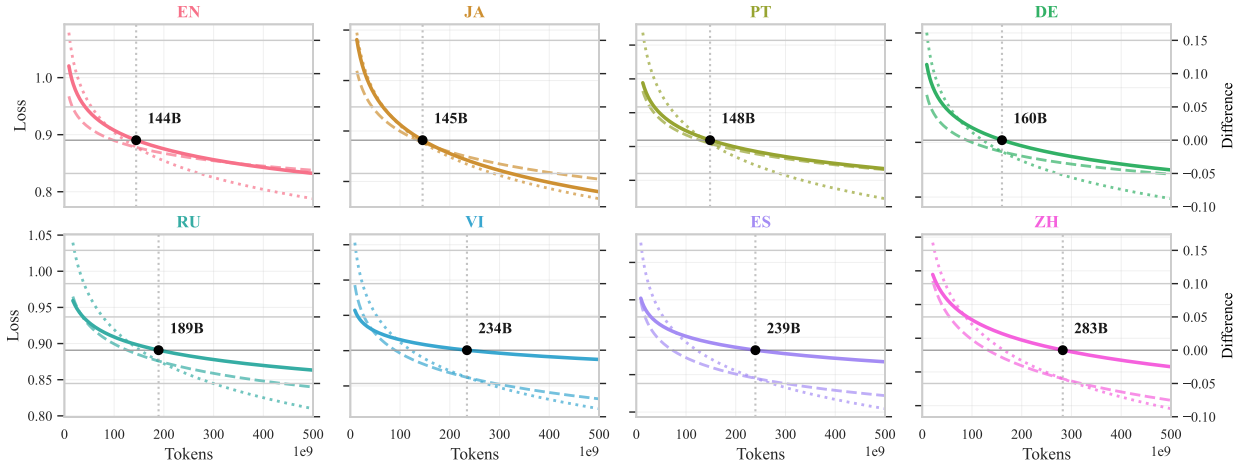


Figure 4.6: For eight languages, we plot the loss curves for 2B parameter models pretraining monolingually from scratch (⋯), finetuning monolingually from the Unimax base model (- - -), and the difference between these losses (—). We annotate the number of tokens at which the pretrained loss becomes better than the finetuning loss. The difference between the interpolated pretraining curve and finetuning curve. When the loss difference reaches zero, pretraining from scratch has surpassed finetuning using equal compute. **Depending on the token (or compute) budget, it is usually more effective to finetune a multilingual checkpoint if there are** $< 144B$ **tokens, and pretrain from scratch if the budget accommodates** $\geq 283B$ **tokens.**

Figure 4.6, we plot the interpolated loss curves for the model pretrained from scratch, the model finetuned from a checkpoint, and the difference between these losses, for several languages. We find that while the warm-started finetuned model performs better at first, pretraining from scratch eventually surpasses its performance after $144B - 283B$ tokens, depending on the language. Note that we leave the reason for these convergence differences to future work, without a clear hypothesis. Though English pretraining converges the fastest, it is also allocated a 5% sampling rate within Unimax, whereas each of the other languages are allocated 1.4%.

From the number of training tokens D we can also infer the inflection point which decides whether it is more computationally efficient ($C = 6ND$) to pretrain from scratch or finetune from the multilingual checkpoint. Using our range of model size experiments, we can estimate

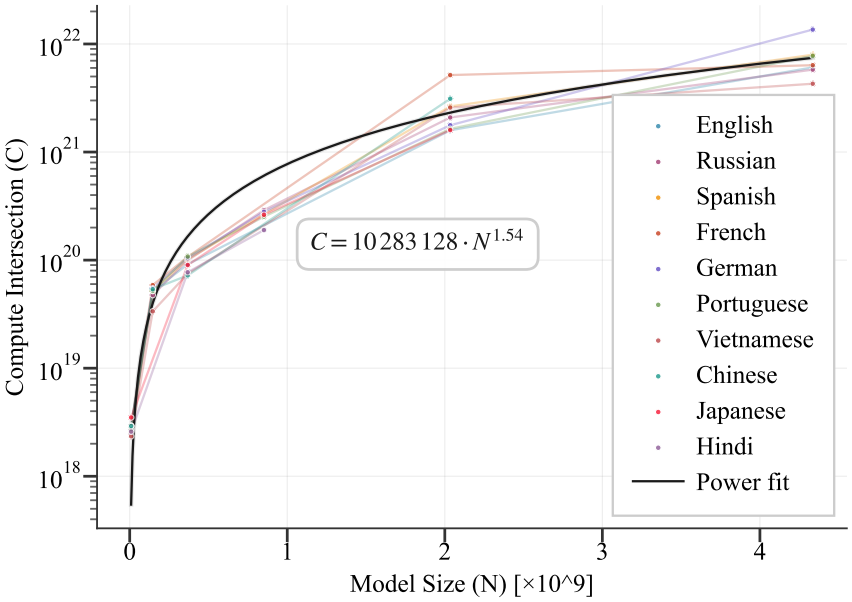


Figure 4.7: In terms of N, we model the compute budget C required for a model pretrained from scratch to outperform the model finetuned from the generic Unimax multilingual checkpoint. **We estimate this relationship as** $C = 10283128 \times N^{1.65}$**.**

a best fit power law in Figure 4.7, relating N to C. We find $log(C) = 1113708 \times N^{1.65}$ to hold consistently across languages. For simplicity, we choose not to account for data repetition in this model. Practitioners may use this to estimate whether pretraining or finetuning will yield greater performance for their compute budget. Note one limitation to consider: not all multilingual base models are trained with the same mixture or for sufficiently long—and these factors would impact the pretrain vs finetuning intersection points. We choose a widely used Unimax mixture trained for 1B tokens (Chung et al., 2023). We believe this is a useful heuristic, with reasonable assumptions, to determine the optimal choice between pretraining and finetuning for a given compute budget.

4.1.6 Conclusion

In this work, we answer several scaling questions on training multilingual models, validating our findings on scales up to 8B model parameters. First, we introduce a novel scaling law formulation called Adaptive Transfer Scaling Law (ATLAS) that significantly improves scaling law fit by more than 0.3 R^2 for multilingual mixtures. To aid our analysis, we also derive transfer scores for 38 langauages. We find that English is often a strong positive transfer language, alongside languages with shared script and language families, though we note that this transfer is often asymmetric. However, there is still a compute-efficiency tax when we train multilingual models. We quantify this "curse of multilinguality" and provide a mathematical trade-off between model capacity and language coverage: we find that the performance penalty from adding languages is mitigated more effectively by scaling model size (N) rather than training tokens (D). Finally, we quantitatively answer a critical practical question for practitioners: determining the compute crossover point between pretraining and finetuning. We find that depending on the language, for a budget under 144-283B tokens, finetuning a general-purpose multilingual checkpoint is more efficient. Collectively, these contributions provide valuable guidelines for practitioners determining data mixtures, language coverage and computational requirements for creating multilingual foundation models.

Chapter 5
CONCLUSION

Throughout this dissertation, I argue that qualitative and quantitative measurement are critical for improving foundation models across both data collection and model training.

Data Collection In Chapter 2, I document the creation life-cycle of `MADLAD-400` and show how qualitative audits surface failure modes that automatic filters miss. Through self-audits, we uncover systematic issues in web data such as mislabeled language codes, domain skew toward religious texts, and script-level noise that are not caught by applying quantitative heuristics (Raffel et al., 2020) or automated filters (Caswell et al., 2020). We use multiple rounds of audits to iteratively improve our filters, through which we see both improved datasets and models.

Model Training In Chapter 3, through a survey of over 50 scaling law papers we show how most scaling law papers do not share sufficient information to reproduce the papers' conclusions. We empirically show how key experimental design choices such as the form of the power law, training hyperparameters, compute measures, and the optimization procedure can change dramatically change the study results by a magnitude of over 10^2 times.

Data Optimality In Chapter 4, I derive scaling laws for multilingual pretraining at scale. I propose ATLAS, a language-agnostic scaling law that improves out-of-sample R^2 by more than 0.3 over prior work. I provide practictioners with tools to make quantitative decisions about expanding language coverages in a compute efficient manner across both pre-training and mid-training.

5.1 Future Work

The contributions of this dissertation have focused primarily on *pretrained foundation models*. However, foundation models are increasingly heterogeneous systems: they are shaped by not

only pretraining, but also by posttraining and inference -time scaffolding. They are also trained on increasingly heterogeneous data distributions, enabling them to integrate information across modalities, and to act through agentic systems to complete complex tasks on behalf of users. Therefore, the quantitative questions that are most critical for guiding training decisions are likely to change in several important ways. I discuss these directions below.

Beyond pretraining data mixtures This dissertation focuses on pretraining, but modern pipelines also involve mid-training and post-training alignment stages. Developing scaling laws and quantitative tools for each of these stages (Khatri et al., 2025), and studying the interplay of these stages with each other at scale (Zhang et al., 2025), remains an open problem. This challenge is compounded by how each stage is trained on heterogeneous data, introducing additional axes along which data mixture, training objective, and model behavior may interact.

Measuring what matters Throughout this work and in most scaling law studies (Li et al., 2025a), test loss is the primary metric for evaluating scaling behavior. However, test loss and downstream task performance on a specific task of interest do not necessarily correlate with each other Schaeffer et al. (2023). Developing tools for predicting downstream task performance, and desirable agent behaviors and core model capabilities at scale is an important open problem.

Models that know what to learn from The contributions in this dissertation treat data curation as a preprocessing step that is finalized before the final training run. However, future training pipelines may increasingly involve data distributions that change over the course of training, especially as challenging tasks on the frontier motivate inference-time compute scaling and continual learning. In such settings, models that can identify which examples (Jiang et al., 2025), or even which parts of examples (Von Oswald et al., 2023), are worth learning from would require shifting the role of data curation from a fixed input of training into an adaptive component of the learning algorithm.

BIBLIOGRAPHY

StatMT. https://www.statmt.org/. Accessed: 2022-05-03.

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. Towards a cleaner document-oriented multilingual crawled corpus. *arXiv preprint arXiv:2201.06642*, 2022.

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

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pp. 265–279. PMLR, 2023.

Željko Agic and Ivan Vulic. Jw300: A wide-coverage parallel corpus for low-resource languages. Association for Computational Linguistics, 2019.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*, 2023.

Ibrahim M Alabdulmohsin, Behnam Neyshabur, and Xiaohua Zhai. Revisiting neural scaling laws in language and vision. *Advances in Neural Information Processing Systems*, 35:22300–22312, 2022.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*, 2024.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pp. 173–182. PMLR, 2016.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

Anthropic. System card: Claude opus 4 & claude sonnet 4. Technical report, Anthropic, May 2025. URL https://www-cdn.anthropic.com/4263b940cabb546aa0e3283f35b686f4f3b2ff47.pdf.

Newsha Ardalani, Carole-Jean Wu, Zeliang Chen, Bhargav Bhushanam, and Adnan Aziz. Understanding scaling laws for recommendation models. *arXiv preprint arXiv:2208.08489*, 2022.

Catherine Arnett and Benjamin Bergen. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6607–6623, 2025.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*, 2023.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27), June 2024. ISSN 1091-6490. doi: 10.1073/pnas.2311878121. URL http://dx.doi.org/10.1073/pnas.2311878121.

Jack Bandy and Nicholas Vincent. Addressing" documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*, 2021.

Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pp. 26–33, 2001.

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, and Orhan Firat. Data scaling laws in nmt: The effect of noise and architecture. In *International Conference on Machine Learning*, pp. 1466–1482. PMLR, 2022.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. Building Machine Translation Systems for the Next Thousand Languages. *arXiv e-prints*, art. arXiv:2205.03983, May 2022.

Christos Baziotis, Biao Zhang, Alexandra Birch, and Barry Haddow. When does monolingual data help multilingual translation: The role of domain and model scale. *arXiv preprint arXiv:2305.14124*, 2023.

Tamay Besiroglu, Ege Erdil, Matthew Barnett, and Josh You. Chinchilla scaling: A replication attempt. *arXiv preprint arXiv:2404.10102*, 2024.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model, 2022. URL https://arxiv.org/abs/2204.06745.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

David Brandfonbrener, Nikhil Anand, Nikhil Vyas, Eran Malach, and Sham Kakade. Loss-to-loss prediction: Scaling laws for all datasets. *arXiv preprint arXiv:2411.12925*, 2024.

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

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. Broken neural scaling laws. *arXiv preprint arXiv:2210.14891*, 2022.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus, 2020. URL `https://arxiv.org/abs/2010.14571`.

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

Yangyi Chen, Binxuan Huang, Yifan Gao, Zhengyang Wang, Jingfeng Yang, and Heng Ji. Scaling laws for predicting downstream performance in llms. *arXiv preprint arXiv:2410.08527*, 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, et al. Palm: Scaling language modeling with pathways, 2022.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*, 2023.

Aidan Clark, Diego de Las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. In *International conference on machine learning*, pp. 4057–4086. PMLR, 2022.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Ian Covert, Wenlong Ji, Tatsunori Hashimoto, and James Zou. Scaling laws for the value of individual data points in machine learning. *arXiv preprint arXiv:2405.20456*, 2024.

DeepSeek-AI. Deepseek-v3 technical report, 2024. URL `https://arxiv.org/abs/2412.19437`.

Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning*, pp. 7750–7774. PMLR, 2023.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, 2022.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*, 2021.

Jasha Droppo and Oguz Elibol. Scaling laws for acoustic models. *arXiv preprint arXiv:2106.09488*, 2021.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Miquel Esplà-Gomis, Mikel L Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. Paracrawl: Web-scale parallel corpora for the languages of the eu. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pp. 118–119, 2019.

Christian Federmann, Tom Kocmi, and Ying Xin. Ntrex-128–news test references for mt evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pp. 21–24, 2022.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1): 5232–5270, 2022.

Patrick Fernandes, Behrooz Ghorbani, Xavier Garcia, Markus Freitag, and Orhan Firat. Scaling laws for multilingual neural machine translation. *arXiv preprint arXiv:2302.09650*, 2023.

Aloka Fernando, Surangika Ranathunga, and Gihan Dias. Data augmentation and terminology integration for domain-specific sinhala-english-tamil statistical machine translation. *arXiv preprint arXiv:2011.02821*, 2020.

Matthew J Filipovich, Alessandro Cappelli, Daniel Hesslow, and Julien Launay. Scaling laws beyond backpropagation. *arXiv preprint arXiv:2210.14593*, 2022.

Elias Frantar, Carlos Riquelme, Neil Houlsby, Dan Alistarh, and Utku Evci. Scaling laws for sparsely-connected foundation models. *arXiv preprint arXiv:2309.08520*, 2023.

Markus Freitag and Orhan Firat. Complete multilingual neural machine translation. *CoRR*, abs/2010.10239, 2020. URL https://arxiv.org/abs/2010.10239.

Jay Gala, Pranjal A Chitale, Raghavan AK, Sumanth Doddapaneni, Varun Gumma, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*, 2023.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. The unreasonable effectiveness of few-shot learning for machine translation. *arXiv preprint arXiv:2302.01398*, 2023.

Ce Ge, Zhijian Ma, Daoyuan Chen, Yaliang Li, and Bolin Ding. Bimix: A bivariate data mixing law for language model pretraining. *arXiv preprint arXiv:2405.14908*, 2024a.

Ce Ge, Zhijian Ma, Daoyuan Chen, Yaliang Li, and Bolin Ding. Data mixing made efficient: A bivariate scaling law for language model pretraining. *arXiv preprint arXiv:2405.14908*, 2024b.

Jonas Geiping, Micah Goldblum, Gowthami Somepalli, Ravid Shwartz-Ziv, Tom Goldstein, and Andrew Gordon Wilson. How much data are augmentations worth? an investigation into scaling laws, invariance, and implicit regularization. *arXiv preprint arXiv:2210.06441*, 2022.

Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. Scaling laws for neural machine translation. *arXiv preprint arXiv:2109.07740*, 2021.

Michel L Goldstein, Steven A Morris, and Gary G Yen. Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 41: 255–258, 2004.

Google DeepMind. Gemini 2.5 deep think model card. Technical report, Google DeepMind, August 2025. URL `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Deep-Think-Model-Card.pdf`. Model card published/updated August 1, 2025.

Mitchell A Gordon, Kevin Duh, and Jared Kaplan. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5915–5922, 2021.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. 2021.

Sachin Goyal, Pratyush Maini, Zachary C Lipton, Aditi Raghunathan, and J Zico Kolter. Scaling laws for data filtering–data curation cannot be compute agnostic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22702–22711, 2024.

Hendrik Johannes Groenewald and Wildrich Fourie. Introducing the autshumato integrated translation environment. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*, 2009.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Barry Haddow and Faheem Kirefu. Pmindia–a collection of parallel corpora of languages of india. *arXiv preprint arXiv:2001.09907*, 2020.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. *arXiv preprint arXiv:2405.18392*, 2024.

Tatsunori Hashimoto. Model performance scaling with multiple data sources. In *International Conference on Machine Learning*, pp. 4107–4116. PMLR, 2021.

Yifei He, Alon Benhaim, Barun Patra, Praneetha Vaddamanu, Sanchit Ahuja, Parul Chopra, Vishrav Chaudhary, Han Zhao, and Xia Song. Scaling laws for multilingual language models. *arXiv preprint arXiv:2410.12883*, 2024.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Jacob Hilton, Jie Tang, and John Schulman. Scaling laws for single-agent reinforcement learning. *arXiv preprint arXiv:2301.13442*, 2023.

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, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Adam Gleave, Laurent Sifre, Geoffrey Irving, Shane Legg, and Demis Hassabis. Training compute-optimal large language models, 2022.

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

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518. Springer, 1992.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André FT Martins, François Yvon, et al. Glot500: Scaling multilingual corpora and language models to 500 languages. *arXiv preprint arXiv:2305.12182*, 2023.

Maor Ivgi, Yair Carmon, and Jonathan Berant. Scaling laws under the microscope: Predicting transformer performance from small scale experiments. *arXiv preprint arXiv:2202.06387*, 2022.

Yiding Jiang, Allan Zhou, Zhili Feng, Sadhika Malladi, and Zico Kolter. Adaptive data optimization: Dynamic sample selection with scaling laws. In *International Conference on Learning Representations*, volume 2025, pp. 58014–58037, 2025.

Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. The nunavut hansard inuktitut–english parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 2562–2572, 2020.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351, 2017.

Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. Bilex rx: Lexical data augmentation for massively multilingual machine translation, 2023.

Andy L Jones. Scaling scaling laws with board games. *arXiv preprint arXiv:2104.03113*, 2021.

Feiyang Kang, Yifan Sun, Bingbing Wen, Si Chen, Dawn Song, Rafid Mahmood, and Ruoxi Jia. Autoscale: Scale-aware data mixing for pre-training llms. *arXiv preprint arXiv:2407.20177*, 2024.

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

Aditya Armaan Khan, Mason Stephen Shipton, David Anugraha, Kaiyao Duan, Phuong H Hoang, Eric Khiu, A Seza Doğruöz, and Annie Lee. Uriel+: Enhancing linguistic inclusion and usability

in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6937–6952, 2025.

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for llms. *arXiv preprint arXiv:2510.13786*, 2025.

Young Jin Kim, Ammar Ahmad Awan, Alexandre Muzio, Andres Felipe Cruz Salinas, Liyang Lu, Amr Hendy, Samyam Rajbhandari, Yuxiong He, and Hany Hassan Awadalla. Scalable and efficient moe training for multitask multilingual models. *arXiv preprint arXiv:2109.10465*, 2021.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pp. 79–86, 2005.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan Van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022.

T Kudo. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018.

Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. Beyond distillation: Task-level mixture-of-experts for efficient inference. *arXiv preprint arXiv:2110.03742*, 2021.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. Llms beyond english: Scaling the multilingual capability of llms with cross-lingual feedback. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 8186–8213, 2024.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826, 2022.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 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*, pp. 8608–8621, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.590.

Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. (mis) fitting: A survey of scaling laws. *arXiv preprint arXiv:2502.18969*, 2025a.

Margaret Li, Sneha Kudugunta, and Luke Zettlemoyer. (mis)fitting: A survey of scaling laws, 2025b.

Zihao Li, Shaoxiong Ji, Hengyu Luo, and Jörg Tiedemann. Rethinking multilingual continual pretraining: Data mixing for adapting llms across languages and resources. *arXiv preprint arXiv:2504.04152*, 2025c.

Daniel Liebling, Katherine Heller, Samantha Robertson, and Wesley Deng. Opportunities for human-centered evaluation of machine translation systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 229–240, 2022.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle

Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*, 2021.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. *arXiv preprint arXiv:2407.01492*, 2024.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, and David Mimno. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*, 2023a.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Alexandra Sasha Luccioni and Joseph D Viviano. What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. *arXiv preprint arXiv:2105.02732*, 2021.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*, 2018.

AI Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *https://ai. meta. com/blog/llama-4-multimodal-intelligence/, checked on*, 4(7):2025, 2025.

Hiroaki Mikami, Kenji Fukumizu, Shogo Murai, Shuji Suzuki, Yuta Kikuchi, Taiji Suzuki, Shin-ichi Maeda, and Kohei Hayashi. A scaling law for syn-to-real transfer: How much is your pre-training effective?

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. Aspec: Asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2204–2208, 2016.

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

Oren Neumann and Claudius Gros. Scaling laws for a multi-agent reinforcement learning model. *arXiv preprint arXiv:2210.00849*, 2022.

NLLBTeam, 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. No language left behind: Scaling human-centered machine translation. 2022.

OpenAI. Gpt-5 system card. Technical report, OpenAI, August 2025. URL `https://cdn.openai.com/gpt-5-system-card.pdf`.

OpenAI. Introducing GPT-5. OpenAI Blog, August 2025. URL `https://openai.com/index/introducing-gpt-5/`. Accessed: 2025-11-03.

R OpenAI. Gpt-4 technical report. *arXiv*, pp. 2303–08774, 2023.

Gabriel Orlanski, Kefan Xiao, Xavier Garcia, Jeffrey Hui, Joshua Howland, Jonathan Malmaud, Jacob Austin, Rishah Singh, and Michele Catasta. Measuring the impact of programming language distribution. *arXiv preprint arXiv:2302.01973*, 2023.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

Tim Pearce and Jinyeop Song. Reconciling kaplan and chinchilla scaling laws. *arXiv preprint arXiv:2406.12907*, 2024a.

Tim Pearce and Jinyeop Song. Reconciling kaplan and chinchilla scaling laws, 2024b. URL https://arxiv.org/abs/2406.12907.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023. URL https://arxiv.org/abs/2306.01116.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Jerin Philip, Vinay P Namboodiri, and CV Jawahar. A baseline neural machine translation system for indian languages. *arXiv preprint arXiv:1907.12437*, 2019.

Michael Poli, Armin W Thomas, Eric Nguyen, Pragaash Ponnusamy, Björn Deiseroth, Kristian Kersting, Taiji Suzuki, Brian Hie, Stefano Ermon, Christopher Ré, et al. Mechanistic design and scaling of hybrid architectures. *arXiv preprint arXiv:2403.17844*, 2024.

Tomer Porian, Mitchell Wortsman, Jenia Jitsev, Ludwig Schmidt, and Yair Carmon. Resolving discrepancies in compute-optimal scaling of language models. *arXiv preprint arXiv:2406.19146*, 2024.

Matt Post. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*, 2018.

Gabriele Prato, Simon Guiroy, Ethan Caballero, Irina Rish, and Sarath Chandar. Scaling laws for the few-shot adaptation of pre-trained image classifiers. *arXiv preprint arXiv:2110.06990*, 2021.

Vitaly Protasov, Elisei Stakovskii, Ekaterina Voloshina, Tatiana Shavrina, and Alexander Panchenko. Super donors and super recipients: Studying cross-lingual transfer between high-resource and low-resource languages. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pp. 94–108, 2024.

Reid Pryzant, Yongjoo Chung, Dan Jurafsky, and Denny Britz. Jesc: Japanese-english subtitle corpus. *arXiv preprint arXiv:1710.10639*, 2017.

Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yinghao Ma, Feiyu Duan, Zhiqi Bai, Jiakai Wang, Yuanxing Zhang, et al. D-cpt law: Domain-specific continual pre-training scaling law for large language models. *Advances in Neural Information Processing Systems*, 37: 90318–90354, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*, 2019.

Yangjun Ruan, Chris J Maddison, and Tatsunori Hashimoto. Observational scaling laws and the predictability of language model performance. *arXiv preprint arXiv:2405.10938*, 2024.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021.

Nikhil Sardana and Jonathan Frankle. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. *arXiv preprint arXiv:2401.00448*, 2023.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Lucile Saulnier, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, et al. What language model to train if you have one million gpu hours? *arXiv preprint arXiv:2210.15424*, 2022.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? *arXiv preprint arXiv:2304.15004*, 2023.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *arXiv preprint arXiv:1907.05791*, 2019.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pp. 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. Multilingual instruction tuning with just a pinch of multilinguality. *arXiv preprint arXiv:2401.01854*, 2024.

Noam Shazeer. Glu variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Luisa Shimabucoro, Ahmet Ustun, Marzieh Fadaee, and Sebastian Ruder. A post-trainer's guide to multilingual training data: Uncovering cross-lingual transfer dynamics. *arXiv preprint arXiv:2504.16677*, 2025.

Kyuyong Shin, Hanock Kwak, Su Young Kim, Max Nihlén Ramström, Jisu Jeong, Jung-Woo Ha, and Kyung-Min Kim. Scaling law for recommendation models: Towards general-purpose user representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4596–4604, 2023.

Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. *arXiv preprint arXiv:2507.09404*, 2025.

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

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*, 2019.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. Scaling laws with vocabulary: Larger models deserve larger vocabularies. *arXiv preprint arXiv:2407.13623*, 2024.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q Tran, Dani Yogatama, and Donald Metzler. Scaling laws vs model architectures: How does inductive bias influence scaling? *arXiv preprint arXiv:2207.10551*, 2022a.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*, 2022b.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael

Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2024. URL https://arxiv.org/abs/2501.00656.

Jörg Tiedemann. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pp. 2214–2218. Citeseer, 2012.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*, 2022.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Alexander Arno Weber, Klaudia Thellmann, Jan Ebert, Nicolas Flores-Herr, Jens Lehmann, Michael Fromm, and Mehdi Ali. Investigating multilingual instruction-tuning: Do polyglot models demand for multilingual instructions? In *EMNLP*, 2024.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

Mitchell Wortsman, Peter J Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, et al. Small-scale proxies for large-scale transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36, 2024.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Yunhua Zhou, Jun Zhan, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. *arXiv preprint arXiv:2403.16952*, 2024.

Qi Ye, Sachan Devendra, Felix Matthieu, Padmanabhan Sarguna, and Neubig Graham. When and why are pre-trained word embeddings useful for neural machine translation. In *HLT-NAACL*, 2018.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12104–12113, 2022.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*, 2020.

Biao Zhang, Behrooz Ghorbani, Ankur Bapna, Yong Cheng, Xavier Garcia, Jonathan Shen, and Orhan Firat. Examining scaling and transfer of language model architectures for machine translation. In *International Conference on Machine Learning*, pp. 26176–26192. PMLR, 2022.

Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*, 2023.

Charlie Zhang, Graham Neubig, and Xiang Yue. On the interplay of pre-training, mid-training, and rl on reasoning language models. *arXiv preprint arXiv:2512.07783*, 2025.

Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The united nations parallel corpus v1. 0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3530–3534, 2016.

Appendix A

MADLAD-400: SUPPLEMENTARY MATERIALS

A.1 LangID Details

Following *Language Id In the Wild* (Caswell et al., 2020), we trained a Transformer-Base (Vaswani et al., 2017) Semi-Supervised LangId model (SSLID) on 498 languages. The training data is as described in *Language ID in the Wild*, with the differences that 1) training data is sampled to a temperature of `T=3` to reduce over-triggering on low-resource languages; and 2) the data is supplemented with web-crawled data from the same paper (that has already been through the various filters described therein). The purpose of adding this data is to increase robustness to web-domain text, and possibly distill some of the filters used to create the web-crawl. The languages chosen for this model were roughly the top 498 by number of sentences in the dataset reported by *Language ID in the Wild*. The complete list may be seen in Table A.1.

Table A.1: BCP-47 codes, name, script and amount of data associated with all languages in `MADLAD-400`. The last 79 languages, with entries of "-", were languages that the LangID model was trained to detect, but was omitted after the self-audit.

BCP-47	Name	Script	docs (noisy)	docs (clean)	sents (noisy)	sents (clean)	chars (noisy)	chars (clean)
total	-	-	7.8B	4B	148.4B	105.5B	33.3T	18.3T
median	-	-	21.6K	1.5K	202.3K	61.8K	63.6M	8.5M
en	English	Latn	3.6B	1.8B	87.9B	53.4B	15T	9T
ru	Russian	Cyrl	823M	402.5M	823M	12.4B	3.1T	1.8T
es	Spanish	Latn	476.4M	250.9M	8.3B	4.5B	2.1T	1.1T
fr	French	Latn	384.2M	218.9M	7.9B	5B	2T	1T
de	German	Latn	478.6M	225.1M	11.5B	6B	2.2T	1T
it	Italian	Latn	238.9M	126.4M	4.5B	2.5B	1.2T	553.1B
pt	Portuguese	Latn	209.2M	124.2M	4B	2.4B	791.5B	499.8B

pl	Polish	Latn	145.1M	90.9M	3.3B	2.4B	505B	356.4B
nl	Dutch	Latn	134.5M	86.6M	134.5M	2.3B	698.5B	334.5B
vi	Vietnamese	Latn	92.8M	55M	1.6B	1B	342B	228.8B
tr	Turkish	Latn	107M	56.4M	107M	1.2B	328.8B	198.9B
sv	Swedish	Latn	65.2M	35.2M	65.2M	1B	422.6B	153.7B
id	Indonesian	Latn	120.9M	38M	2.2B	747.5M	443B	148.3B
ro	Romanian	Latn	60.8M	35.4M	60.8M	746.4M	244.1B	148.2B
cs	Czech	Latn	72.1M	38.3M	1.7B	1B	272.2B	147.9B
zh	Mandarin Chinese	Hans	29.3M	19.9M	492.3M	298.8M	333B	142.3B
hu	Hungarian	Latn	47.6M	29.7M	1.3B	806.3M	223.6B	134.9B
ja	Japanese	Jpan	23.3M	21.8M	326M	321.6M	133.3B	132.2B
th	Thai	Thai	19M	17.4M	19M	385.8M	118.6B	117.6B
fi	Finnish	Latn	35.8M	20.4M	1B	650.3M	202.2B	101.1B
fa	Persian	Arab	58.1M	23.1M	920.6M	493.5M	220.4B	96.7B
uk	Ukrainian	Cyrl	46.6M	25M	1B	599.9M	164.2B	95.2B
da	Danish	Latn	38.5M	17.9M	1.1B	508M	252B	83.1B
el	Greek	Grek	52.4M	20.9M	808M	445.4M	173.2B	80.9B
no	Norwegian	Latn	34.7M	14.9M	34.7M	498.7M	305.6B	74.8B
bg	Bulgarian	Cyrl	27.2M	12.8M	599.4M	360.3M	95.6B	57.8B
sk	Slovak	Latn	23.2M	11.9M	487.9M	300.6M	77.8B	45.7B
ko	Korean	Kore	19.7M	12.7M	628.6M	471.8M	65.9B	43.8B
ar	Arabic	Arab	67.6M	12.4M	876.6M	182.6M	243B	43.2B
lt	Lithuanian	Latn	15.3M	8.7M	374M	256.9M	58.6B	41.3B
ca	Catalan	Latn	17.9M	9.5M	258.6M	153M	56.5B	34.6B
sl	Slovenian	Latn	12M	6.3M	316M	180M	47.8B	30.5B
he	Hebrew	Hebr	14.1M	7.2M	302.2M	196.8M	54.9B	30.5B
et	Estonian	Latn	8.8M	5.5M	223.8M	176.3M	40.1B	28.7B
lv	Latvian	Latn	8.4M	5M	186.1M	138.5M	36.7B	23.9B
hi	Hindi	Deva	9.9M	4.5M	254.4M	152M	39.9B	20.1B
sq	Albanian	Latn	5.5M	3.6M	5.5M	56.1M	17B	12.7B
ms	Malay	Latn	14.1M	2.3M	14.1M	55.2M	58.8B	12.5B
az	Azerbaijani	Latn	5.2M	3.3M	90.3M	70.9M	16.3B	11.9B
sr	Serbian	Cyrl	4.7M	2M	4.7M	64M	18.6B	11B
ta	Tamil	Taml	5.6M	2.6M	122.5M	81.9M	19.2B	10.6B
hr	Croatian	Latn	23M	2.8M	476.6M	53M	85.1B	9.6B
kk	Kazakh	Cyrl	3.1M	1.8M	87.4M	59.1M	13.4B	8.6B
is	Icelandic	Latn	2.9M	1.6M	73.7M	39.3M	14.9B	6.4B
ml	Malayalam	Mlym	3.7M	2.1M	75M	52M	10.5B	6.3B
mr	Marathi	Deva	2.9M	1.7M	2.9M	50M	8.7B	5.5B
te	Telugu	Telu	2.5M	1.7M	59M	46.4M	7.4B	5.1B
af	Afrikaans	Latn	2.9M	868.7K	51.9M	30M	11.8B	4.8B
gl	Galician	Latn	4.2M	1.3M	45.3M	18.8M	15.6B	4.8B
fil	Filipino	Latn	4.2M	901.5K	67.4M	19.2M	14.6B	4.7B
be	Belarusian	Cyrl	2M	1.1M	48.8M	31.3M	7.2B	4.6B
mk	Macedonian	Cyrl	2.9M	1.4M	41.3M	22.6M	9.1B	4.5B
eu	Basque	Latn	2.1M	1.2M	41.7M	24.8M	6.9B	4.3B
bn	Bengali	Beng	4.3M	1.1M	151.2M	38.6M	16.8B	4.3B
ka	Georgian	Geor	3.1M	936.5K	53.7M	26.6M	10.3B	3.8B
mn	Mongolian	Cyrl	2.2M	879.9K	43.3M	24M	7.9B	3.5B
bs	Bosnian	Cyrl	12.9M	1.4M	163.6M	9M	39.5B	3.3B

uz	Uzbek	Latn	1.4M	669.9K	25.7M	17.5M	5.2B	3.3B
ur	Urdu	Arab	967.2K	467.2K	29M	18.4M	5.2B	2.7B
sw	Swahili	Latn	1.3M	537.8K	1.3M	9.5M	4.6B	2.4B
yue	Cantonese	Hant	465.9K	309.3K	2.8M	2.4M	2.4B	2.3B
ne	Nepali	Deva	876.4K	453.3K	876.4K	20.4M	3.9B	2.2B
kn	Kannada	Knda	1.6M	657.8K	32.9M	19.2M	4.6B	2.2B
kaa	Kara-Kalpak	Cyrl	1.1M	586.4K	19.8M	13.3M	3.8B	2.2B
gu	Gujarati	Gujr	1.3M	659.7K	28.9M	18.1M	3.9B	2.1B
si	Sinhala	Sinh	788K	349.2K	22.1M	16M	3.4B	1.9B
cy	Welsh	Latn	4.9M	430.7K	68.3M	7.4M	26.4B	1.7B
eo	Esperanto	Latn	1.4M	260K	33.9M	9.3M	5.5B	1.7B
la	Latin	Latn	2.9M	319.2K	85.7M	13.8M	8.2B	1.5B
hy	Armenian	Armn	2M	397.5K	31.1M	9.9M	8.1B	1.5B
ky	Kyrghyz	Cyrl	751.1K	367.6K	14.3M	9.6M	2.5B	1.4B
tg	Tajik	Cyrl	789.2K	328.2K	789.2K	7.4M	2.6B	1.4B
ga	Irish	Latn	5.3M	286K	31.7M	6.9M	30.6B	1.4B
mt	Maltese	Latn	1.2M	265.4K	1.2M	5.6M	3.2B	1.3B
my	Myanmar (Burmese)	Mymr	176.5K	172.4K	176.5K	10.1M	1.3B	1.3B
km	Khmer	Khmr	297.8K	285.7K	5M	5M	1.1B	1.1B
tt	Tatar	Cyrl	2.1M	346.9K	60.2M	8.6M	12.1B	1B
so	Somali	Latn	729.2K	293.2K	729.2K	3.1M	2.1B	992.4M
ku	Kurdish (Kurmanji)	Latn	671.9K	218.9K	10.7M	4.9M	2.1B	849.9M
ps	Pashto	Arab	429.9K	252.9K	5.1M	3.6M	1.4B	848.9M
pa	Punjabi	Guru	368.2K	150.6K	368.2K	6M	1.6B	797.1M
rw	Kinyarwanda	Latn	681.8K	226.5K	681.8K	1.9M	1.7B	749.1M
lo	Lao	Laoo	229.1K	216K	2.9M	2.8M	706.9M	697.6M
ha	Hausa	Latn	443.9K	173.5K	4.5M	2.4M	1.3B	630.2M
dv	Dhivehi	Thaa	264.4K	167.2K	4.3M	3.5M	877.3M	603.1M
fy	W. Frisian	Latn	1.7M	210K	12.1M	3.7M	3.7B	592.3M
lb	Luxembourgish	Latn	7.6M	146K	47.1M	3.4M	58.4B	575.5M
ckb	Kurdish (Sorani)	Arab	622.7K	148.9K	5.6M	2.5M	2.2B	572.7M
mg	Malagasy	Latn	295.2K	115.4K	4.5M	2.6M	1.3B	548.5M
gd	Scottish Gaelic	Latn	206K	94.3K	3.7M	2.4M	812M	526M
am	Amharic	Ethi	245.2K	106.3K	7.1M	5.3M	869.9M	509M
ug	Uyghur	Arab	227.1K	106.5K	4.5M	3.1M	998.5M	504.6M
ht	Haitian Creole	Latn	425.6K	110.4K	6.7M	2.6M	994.5M	461.5M
grc	Ancient Greek	Grek	364.8K	70.7K	13.7M	2.8M	2B	417.8M
hmn	Hmong	Latn	241.3K	75.2K	3.5M	1.9M	1.2B	408.8M
sd	Sindhi	Arab	115.6K	65.9K	115.6K	2.4M	561M	380.4M
jv	Javanese	Latn	999.5K	69.5K	13M	2M	2.3B	376.1M
mi	Maori	Latn	711.9K	79.5K	5.9M	1.9M	1.6B	371.9M
tk	Turkmen	Latn	180.2K	82.5K	180.2K	1.8M	575.2M	369M
ceb	Cebuano	Latn	617.5K	66.2K	6.7M	1.6M	1.5B	357.7M
yi	Yiddish	Hebr	160.6K	64.9K	3.3M	1.9M	838.4M	352.6M
ba	Bashkir	Cyrl	372.4K	90.3K	9.3M	2.6M	766.5M	320.7M
fo	Faroese	Latn	382.9K	97.8K	3.9M	1.8M	923.3M	314.9M
or	Odia (Oriya)	Orya	139.6K	100.5K	139.6K	3.1M	437.2M	309.5M
xh	Xhosa	Latn	310.9K	53.7K	2.9M	1.4M	749.5M	287.3M
su	Sundanese	Latn	336.6K	55K	336.6K	1.6M	967.2M	286.7M
kl	Kalaallisut	Latn	85.9K	46K	2.1M	1.5M	403.9M	279.1M

ny	Chichewa	Latn	181.6K	52.2K	181.6K	1.5M	611.2M	277.5M
sm	Samoan	Latn	137.8K	52.6K	1.9M	1.3M	607.9M	276.3M
sn	Shona	Latn	3.1M	60.2K	3.1M	1.2M	10.6B	266M
co	Corsican	Latn	546.7K	55.4K	6.1M	1.3M	1.1B	265.5M
zu	Zulu	Latn	372.3K	53.8K	3.8M	1.2M	1.2B	257.4M
ig	Igbo	Latn	130.4K	54.4K	2.1M	1.4M	846.1M	251.4M
yo	Yoruba	Latn	115K	52.1K	2M	1.2M	415.6M	239M
pap	Papiamento	Latn	259.1K	54.5K	259.1K	1.4M	1.4B	229.9M
st	Sesotho	Latn	96.8K	40.4K	96.8K	1.1M	381.5M	226.9M
haw	Hawaiian	Latn	310.4K	45.7K	7.1M	1M	892M	214.2M
as	Assamese	Beng	53.9K	33.8K	2.4M	1.7M	275.8M	182.1M
oc	Occitan	Latn	2.4M	36.4K	2.4M	1.6M	6.7B	177.6M
cv	Chuvash	Cyrl	599.4K	47.3K	12M	1.6M	1B	168.9M
lus	Mizo	Latn	91.5K	36.4K	1.4M	863.5K	298.3M	167.3M
tet	Tetum	Latn	291K	40.4K	1.9M	475.7K	1.6B	152.3M
gsw	Swiss German	Latn	7.6M	42.7K	64.5M	1M	42.3B	149.2M
sah	Yakut	Cyrl	1.3M	29.2K	1.3M	1.2M	2.2B	148.2M
br	Breton	Latn	705.4K	33.2K	7.8M	731.7K	3.7B	125.4M
rm	Romansh	Latn	238.1K	33.8K	238.1K	603.4K	391M	100.2M
sa	Sanskrit	Deva	154.3K	7.1K	154.3K	1.1M	512.5M	88.8M
bo	Tibetan	Tibt	6.2K	6.2K	1.1M	1.1M	88.7M	88.7M
om	Oromo	Latn	846.1K	18.9K	846.1K	469.8K	1.9B	88.5M
se	N. Sami	Latn	54.3K	23.9K	879.5K	493.3K	148.4M	84.6M
ce	Chechen	Cyrl	59.3K	15K	991.1K	460.1K	130.6M	67.8M
cnh	Hakha Chin	Latn	44.4K	21.6K	688.6K	406.9K	110.8M	63M
ilo	Ilocano	Latn	69.8K	11.8K	889.2K	365.1K	187.9M	59.4M
hil	Hiligaynon	Latn	126.8K	10.6K	1.1M	379.7K	293.5M	57.2M
udm	Udmurt	Cyrl	67.1K	13.4K	942.7K	510.3K	106M	55.5M
os	Ossetian	Cyrl	172.1K	12.6K	172.1K	359.3K	233.5M	50.1M
lg	Luganda	Latn	61.1K	13K	510.9K	166.1K	160.7M	48M
ti	Tigrinya	Ethi	20.8K	7.3K	20.8K	481.3K	95.4M	44.6M
vec	Venetian	Latn	1.1M	11.1K	10M	209.7K	1.8B	43.8M
ts	Tsonga	Latn	34.7K	5.2K	34.7K	248.6K	377.2M	38.8M
tyv	Tuvinian	Cyrl	61.6K	9.1K	596.6K	268.3K	80.2M	38.5M
kbd	Kabardian	Cyrl	154.7K	7.5K	1.4M	257.2K	321.4M	36.8M
ee	Ewe	Latn	14.1K	4.5K	353.6K	246.7K	67.9M	32.8M
iba	Iban	Latn	34K	7.6K	326.9K	126.1K	251.4M	30.5M
av	Avar	Cyrl	107.6K	6.3K	806.1K	190.1K	129M	30.2M
kha	Khasi	Latn	37.8K	12.1K	235.5K	75.2K	88.6M	30.2M
to	Tonga (Tonga Islands)	Latn	14.3K	4.6K	14.3K	149K	58.2M	29.9M
tn	Tswana	Latn	138.2K	4.8K	138.2K	174.4K	302.3M	29.2M
nso	Sepedi	Latn	376.2K	4.4K	376.2K	188.4K	2B	28.2M
fj	Fijian	Latn	17K	4K	410K	164.1K	67.7M	28M
zza	Zaza	Latn	370.1K	6K	3.3M	229.2K	617.3M	26.3M
ak	Twi	Latn	19.5K	4.8K	341.7K	210.2K	74.5M	24.8M
ada	Adangme	Latn	6.5K	3.1K	291.5K	199.2K	38.9M	24.2M
otq	Querétaro Otomi	Latn	17.6K	5.6K	17.6K	114.8K	65M	23.4M
dz	Dzongkha	Tibt	1.9K	1.9K	191.7K	191.7K	22.7M	22.7M
bua	Buryat	Cyrl	9.8K	5.3K	252K	144.6K	38M	21.7M
cfm	Falam Chin	Latn	9.1K	4.9K	199.6K	128.6K	32.9M	21.5M

ln	Lingala	Latn	94.7K	3.3K	718.7K	139K	291.8M	21.5M
chm	Meadow Mari	Cyrl	81.5K	4.7K	929.1K	179.7K	132.2M	21.3M
gn	Guarani	Latn	87.1K	3.9K	770.9K	162.6K	140.7M	20.8M
krc	Karachay-Balkar	Cyrl	359.5K	4.8K	2.3M	153.9K	369.5M	20.7M
wa	Walloon	Latn	70.6K	2.8K	1.5M	127.2K	198.8M	20.4M
hif	Fiji Hindi	Latn	702K	2.4K	7.9M	124.7K	9.1B	19.1M
yua	Yucateco	Latn	10.4K	4K	141.6K	77.6K	36.8M	17.2M
srn	Sranan Tongo	Latn	16.7K	2.3K	16.7K	139.5K	49.1M	17M
war	Waray (Philippines)	Latn	1M	2.9K	114M	96.2K	3.5B	16.1M
rom	Romani	Latn	22.9K	4.2K	22.9K	76.1K	59M	15.9M
bik	Central Bikol	Latn	44.8K	3.1K	376.7K	77K	102.3M	15.7M
pam	Pampanga	Latn	174.2K	2.8K	174.2K	23.3K	324M	15.5M
sg	Sango	Latn	4.2K	2.1K	154K	117.9K	22.6M	15.5M
lu	Luba-Katanga	Latn	10.6K	1.4K	316K	112.1K	54.2M	15.4M
ady	Adyghe	Cyrl	74.9K	4.2K	446.8K	96.9K	67.9M	14.8M
kbp	Kabiyè	Latn	5.9K	3K	247.9K	128.3K	30.8M	14.6M
syr	Syriac	Syrc	3.5K	716	326.4K	197.1K	31.5M	14M
ltg	Latgalian	Latn	13.1K	4.1K	213.7K	87.3K	29.2M	13.9M
myv	Erzya	Cyrl	164.8K	3.1K	164.8K	130K	120.3M	13.8M
iso	Isoko	Latn	3.7K	1.7K	155.8K	111.5K	23M	13.7M
kac	Kachin	Latn	5.9K	2.6K	109.2K	77.4K	26.6M	13.6M
bho	Bhojpuri	Deva	13.6K	4.1K	306.2K	118.5K	37.6M	13.4M
ay	Aymara	Latn	8.1K	2.5K	196.7K	83.8K	34.5M	13.1M
kum	Kumyk	Cyrl	4.2K	2.5K	132.2K	89.7K	18.2M	12.4M
qu	Quechua	Latn	149.7K	2.4K	1M	87K	200.6M	12.2M
za	Zhuang	Latn	824.7K	1.7K	19.2M	53.9K	3B	12.1M
pag	Pangasinan	Latn	49.6K	1.6K	49.6K	88.8K	92.9M	12M
ngu	Guerrero Nahuatl	Latn	3.8K	1.5K	3.8K	87.1K	21.4M	11.8M
ve	Venda	Latn	3.8K	1.9K	97.8K	79.4K	19M	11.7M
pck	Paite Chin	Latn	8.9K	1.3K	8.9K	69.7K	39.8M	11.5M
zap	Zapotec	Latn	5.5K	1.8K	202.3K	93.5K	26.4M	11.4M
tyz	Tày	Latn	8K	1.7K	454.8K	104.6K	46.3M	11.3M
hui	Huli	Latn	2K	1.7K	80.1K	74.7K	11.8M	10.9M
bbc	Batak Toba	Latn	72.3K	1.3K	718.3K	73.2K	151.3M	10.6M
tzo	Tzotzil	Latn	2.8K	1.4K	100.4K	75.7K	15.9M	10.6M
tiv	Tiv	Latn	3.8K	1.1K	3.8K	80.7K	20.4M	10.2M
ksd	Kuanua	Latn	14.9K	2K	533K	78.6K	62.4M	10M
gom	Goan Konkani	Deva	4.6K	2.1K	178.3K	108K	19.8M	10M
min	Minangkabau	Latn	28.2K	1.5K	500.9K	75.6K	70.5M	9.9M
ang	Old English	Latn	66.5K	803	1.8M	86.7K	193M	9.8M
nhe	E. Huasteca Nahuatl	Latn	3K	1.7K	3K	57.7K	15.6M	9.8M
bgp	E. Baluchi	Latn	355.7K	2.4K	5.6M	43.3K	1.1B	9.8M
nzi	Nzima	Latn	2.5K	1.4K	2.5K	71.8K	14.4M	9.4M
nnb	Nande	Latn	4.9K	1.1K	4.9K	70.2K	27.7M	9.1M
nv	Navajo	Latn	17.1K	12.6K	17.1K	86.5K	24.8M	9.1M
zxx	Noise	-	118.8K	1.8K	3.8M	49.3K	501K	6.6K
bci	Baoulé	Latn	7.4K	1.3K	124.8K	87.1K	32.8M	9M
kv	Komi	Cyrl	59.1K	1.9K	584.3K	88.8K	91.4M	9M
new	Newari	Deva	6.6K	1.6K	6.6K	85K	21.2M	8.8M
mps	Dadibi	Latn	2.7K	1.2K	132.8K	71.9K	16M	8.7M

alt	S. Altai	Cyrl	2.6K	1.4K	110.1K	65.9K	14.3M	8.7M
meu	Motu	Latn	5.9K	1.7K	232.1K	72.6K	27.2M	8.6M
bew	Betawi	Latn	311.1K	2.7K	10.4M	58.4K	1.4B	8.5M
fon	Fon	Latn	5.3K	1.1K	222.9K	67.3K	34M	8.3M
iu	Inuktitut	Cans	5.4K	2.5K	92.6K	53.1K	17.5M	8.3M
abt	Ambulas	Latn	1.6K	1.3K	122.7K	110.3K	9.6M	8.2M
mgh	Makhuwa-Meetto	Latn	5.5K	1.2K	151.8K	61.2K	24.1M	8.2M
mnw	Mon	Mymr	1.1K	1.1K	144.8K	144.7K	8.1M	8.1M
tvl	Tuvalu	Latn	2.3K	933	72.9K	53.6K	12.6M	8.1M
dov	Dombe	Latn	3.5K	923	129.8K	56.7K	20.7M	8M
tlh	Klingon	Latn	516.9K	3.1K	516.9K	46.9K	1.4B	7.8M
ho	Hiri Motu	Latn	2K	1.5K	57K	47.8K	12.3M	7.8M
kw	Cornish	Latn	176.9K	2.3K	1M	51.6K	327.8M	7.7M
mrj	Hill Mari	Cyrl	97.1K	1.4K	97.1K	60.3K	100.6M	7.6M
meo	Kedah Malay	Latn	790.7K	4.7K	16.5M	39K	3B	7.5M
crh	Crimean Tatar	Cyrl	5.1K	1.2K	170.9K	61.8K	18.8M	7.5M
mbt	Matigsalug Manobo	Latn	1.6K	969	86K	45.4K	14.6M	7.5M
emp	N. Emberá	Latn	3.6K	1.2K	106.4K	75.4K	14.5M	7.4M
ace	Achinese	Latn	65.5K	966	632.5K	32.5K	146.1M	7.4M
ium	Iu Mien	Latn	100.3K	1.7K	6.2M	54.9K	314M	7.4M
mam	Mam	Latn	23K	1.5K	446.3K	52.9K	70.4M	7.2M
gym	Ngäbere	Latn	1.5K	820	73.7K	49.6K	10.3M	6.9M
mai	Maithili	Deva	54.3K	1.2K	1M	60.2K	156M	6.8M
crs	Seselwa Creole French	Latn	7.6K	873	282.4K	40.1K	40.1M	6.8M
pon	Pohnpeian	Latn	5.7K	1.5K	167.8K	48.7K	18.3M	6.7M
ubu	Umbu-Ungu	Latn	2.2K	846	113.5K	47.5K	15.9M	6.7M
fip	Fipa	Latn	3.7K	729	165.6K	49K	25.7M	6.6M
quc	K’iche’	Latn	4.4K	1.5K	89.2K	41.2K	16.6M	6.4M
gv	Manx	Latn	501.9K	1.6K	18.8M	26.9K	933.1M	6.2M
kj	Kuanyama	Latn	112.2K	2.1K	881.8K	22.6K	339.6M	6M
btx	Batak Karo	Latn	3.1K	1K	81.7K	43.9K	13.1M	5.9M
ape	Bukiyip	Latn	7K	814	147K	56.1K	71M	5.8M
chk	Chuukese	Latn	2.8K	1.1K	98.8K	44K	12M	5.8M
rcf	Réunion Creole French	Latn	21.6K	2.6K	21.6K	50.5K	30.2M	5.7M
shn	Shan	Mymr	889	788	46.4K	46.2K	5.7M	5.7M
tzh	Tzeltal	Latn	1.7K	702	41.7K	33.9K	9.3M	5.6M
mdf	Moksha	Cyrl	71K	1.6K	394.7K	45.1K	65.8M	5.5M
ppk	Uma	Latn	2.6K	1.1K	85.8K	34.9K	13.2M	5.5M
ss	Swati	Latn	8.1K	1.1K	8.1K	30.4K	23.7M	5.5M
gag	Gagauz	Latn	33.9K	1.6K	491K	37K	84.9M	5.2M
cab	Garifuna	Latn	1.2K	629	50.4K	37.5K	7.5M	5.1M
kri	Krio	Latn	39.1K	786	271.2K	38.8K	86.4M	5M
seh	Sena	Latn	5.6K	545	68.8K	37.2K	14.9M	4.9M
ibb	Ibibio	Latn	74.1K	818	516.5K	36.3K	190.9M	4.9M
tbz	Ditammari	Latn	5.1K	1.1K	128.7K	37.5K	22M	4.8M
bru	E. Bru	Latn	3K	1.1K	89.7K	48.2K	12.9M	4.8M
enq	Enga	Latn	7.1K	793	241.9K	39.1K	68.5M	4.8M
ach	Acoli	Latn	2K	915	63K	40.1K	9M	4.7M
cuk	San Blas Kuna	Latn	4.1K	899	76.5K	34.3K	24.7M	4.6M
kmb	Kimbundu	Latn	1.3K	538	60.4K	36.9K	8.4M	4.6M

wo	Wolof	Latn	36.4K	871	303.4K	25.4K	213.4M	4.5M
kek	Kekchí	Latn	3.2K	782	70.4K	38.4K	13.6M	4.4M
qub	Huallaga Huánuco Quechua	Latn	972	705	61K	51.1K	5.9M	4.4M
tab	Tabassaran	Cyrl	7.8K	1.2K	226.4K	26.8K	33.7M	4.4M
bts	Batak Simalungun	Latn	3.2K	869	109.1K	29.1K	20.8M	4.2M
kos	Kosraean	Latn	2.2K	881	44.6K	27.8K	6.5M	4.2M
rwo	Rawa	Latn	938	572	938	45.5K	5.1M	4.2M
cak	Kaqchikel	Latn	1.2K	617	70.4K	32.6K	7.6M	4.2M
tuc	Mutu	Latn	3.5K	635	193.2K	50.3K	17.2M	4.1M
bum	Bulu	Latn	4.7K	559	103.8K	36.5K	18.8M	4M
cjk	Chokwe	Latn	3.6K	586	144.1K	24.1K	22.5M	3.9M
gil	Gilbertese	Latn	3.9K	586	151.5K	24.1K	24.1M	3.9M
stq	Saterfriesisch	Latn	111.9K	809	111.9K	27.7K	243.1M	3.8M
tsg	Tausug	Latn	353.8K	789	353.8K	17.9K	1.1B	3.8M
quh	S. Bolivian Quechua	Latn	1K	501	42K	29.9K	5.8M	3.7M
mak	Makasar	Latn	1K	555	32.5K	20.4K	6.1M	3.7M
arn	Mapudungun	Latn	2.4K	593	64.5K	26.2K	10.2M	3.7M
ban	Balinese	Latn	8K	637	150.9K	16.3K	35.4M	3.6M
jiv	Shuar	Latn	1.7K	696	80.9K	32K	9.6M	3.5M
sja	Epena	Latn	1.3K	527	67.7K	24.9K	7.7M	3.4M
yap	Yapese	Latn	1.9K	638	37.6K	19.5K	6.9M	3.3M
tcy	Tulu	Knda	10.7K	632	338.7K	37.1K	41.6M	3.3M
toj	Tojolabal	Latn	736	452	736	26.1K	4.3M	3.3M
twu	Termanu	Latn	2.5K	539	109.9K	24.4K	14.2M	3.2M
xal	Kalmyk	Cyrl	71.8K	913	498.5K	30.8K	64.7M	3.2M
amu	Guerrero Amuzgo	Latn	1.8K	511	72K	25.2K	9.6M	3.2M
rmc	Carpathian Romani	Latn	2.4K	738	2.4K	25.8K	7.9M	3.2M
hus	Huastec	Latn	825	569	26.5K	23.7K	4.4M	3.1M
nia	Nias	Latn	2K	408	2K	25K	11.3M	3.1M
kjh	Khakas	Cyrl	1.5K	672	42.8K	28.7K	4.5M	3.1M
bm	Bambara	Latn	21.9K	702	172.3K	24.5K	48.4M	3M
guh	Guahibo	Latn	1.9K	331	104.9K	28.4K	11.2M	3M
mas	Masai	Latn	15.2K	405	216.8K	17.6K	42.1M	3M
acf	St Lucian Creole French	Latn	4.9K	730	81.9K	24.6K	11.6M	3M
dtp	Kadazan Dusun	Latn	4.6K	1.3K	51.2K	7.9K	12.7M	3M
ksw	S'gaw Karen	Mymr	560	536	16.1K	16K	2.9M	2.9M
bzj	Belize Kriol English	Latn	983	404	33.6K	26.4K	4.5M	2.9M
din	Dinka	Latn	128.4K	611	885.8K	23.6K	210M	2.9M
zne	Zande	Latn	1.3K	239	61.9K	21.3K	8.2M	2.8M
mad	Madurese	Latn	103.8K	509	500.6K	18.5K	111.8M	2.8M
msi	Sabah Malay	Latn	686.7K	1.9K	686.7K	22.6K	2.6B	2.7M
mag	Magahi	Deva	631	138	62.6K	22.1K	10.7M	2.6M
mkn	Kupang Malay	Latn	956	402	33.1K	25.4K	3.4M	2.6M
kg	Kongo	Latn	4.7K	365	85.5K	21.7K	16.6M	2.6M
lhu	Lahu	Latn	46K	377	975K	15.7K	208.6M	2.5M
ch	Chamorro	Latn	12.9K	449	147.5K	16K	63.5M	2.5M
qvi	Imbabura H. Quichua	Latn	1.2K	266	48.4K	19.3K	6.5M	2.3M
mh	Marshallese	Latn	4.6K	296	235.1K	13K	24.9M	2.2M
djk	E. Maroon Creole	Latn	560	246	30.9K	24.4K	3.7M	2.2M
sus	Susu	Latn	664	437	664	15.2K	3.7M	2.1M

mfe	Morisien	Latn	7.5K	320	198.8K	18.2K	26.9M	2.1M
srm	Saramaccan	Latn	847	227	847	17.3K	6.3M	2M
dyu	Dyula	Latn	1.2K	483	55.8K	19.7K	5.7M	2M
ctu	Chol	Latn	690	366	35.5K	20.6K	3.6M	2M
gui	E. Bolivian Guaraní	Latn	1.1K	409	62.7K	24.8K	6.5M	2M
pau	Palauan	Latn	1.7K	185	1.7K	13.1K	12.4M	2M
inb	Inga	Latn	387	343	17.3K	17K	2M	1.9M
bi	Bislama	Latn	71.9K	311	308.5K	13.6K	132.4M	1.9M
mni	Meiteilon (Manipuri)	Beng	1.2K	290	38.1K	13.2K	6.4M	1.8M
guc	Wayuu	Latn	537	214	22.9K	12.5K	3.4M	1.8M
jam	Jamaican Creole English	Latn	12.7K	416	68.5K	15.8K	25.8M	1.7M
wal	Wolaytta	Latn	2.6K	286	128K	14K	17M	1.7M
jac	Popti’	Latn	8.2K	303	61.6K	11.9K	15.7M	1.7M
bas	Basa (Cameroon)	Latn	4.2K	216	105.2K	14.9K	25.7M	1.7M
gor	Gorontalo	Latn	1.7K	303	53.3K	6.5K	9.4M	1.7M
skr	Saraiki	Arab	3.8K	107	279.3K	17.1K	32.2M	1.7M
nyu	Nyungwe	Latn	1.2K	195	1.2K	11K	7.7M	1.6M
noa	Woun Meu	Latn	902	234	902	11.5K	5.2M	1.6M
sda	Toraja-Sa’dan	Latn	1.6K	317	43.2K	6.2K	15.8M	1.6M
gub	Guajajára	Latn	31.7K	271	160.4K	25K	44.7M	1.6M
nog	Nogai	Cyrl	970	419	970	11K	2.6M	1.6M
cni	Asháninka	Latn	1K	261	46K	14K	5.9M	1.6M
teo	Teso	Latn	2.8K	274	131.5K	13.7K	15.3M	1.6M
tdx	Tandroy-Mahafaly Malagasy	Latn	1.7K	262	26.3K	13.2K	7M	1.6M
sxn	Sangir	Latn	587	197	587	9.9K	3.4M	1.5M
rki	Rakhine	Mymr	331	251	331	7.8K	1.6M	1.5M
nr	South Ndebele	Latn	10.7K	246	10.7K	11.3K	49M	1.5M
frp	Arpitan	Latn	148K	550	3.5M	8.2K	535.4M	1.4M
alz	Alur	Latn	2.2K	195	59.3K	12.2K	7.9M	1.4M
taj	E. Tamang	Deva	146	65	21.6K	14.3K	2.3M	1.4M
lrc	N. Luri	Arab	42.4K	587	351.9K	9K	85.3M	1.4M
cce	Chopi	Latn	847	116	23.2K	11K	3.3M	1.3M
rn	Rundi	Latn	8.2K	323	8.2K	11.1K	33.2M	1.3M
jvn	Caribbean Javanese	Latn	1K	213	36.2K	7.8K	5.3M	1.2M
hvn	Sabu	Latn	737	200	33.9K	7K	4.3M	1.2M
nij	Ngaju	Latn	1K	183	1K	9.2K	4.7M	1.2M
dwr	Dawro	Latn	452	215	22.1K	11.1K	2.2M	1.2M
izz	Izii	Latn	423	237	21.7K	14.5K	2.1M	1.1M
msm	Agusan Manobo	Latn	520	177	520	8.6K	2.5M	1.1M
bus	Bokobaru	Latn	467	322	21.4K	12.1K	2.1M	1.1M
ktu	Kituba (DRC)	Latn	3.3K	144	115.5K	7.8K	18.5M	1.1M
chr	Cherokee	Cher	964	301	33.8K	7.5K	4.7M	1M
maz	Central Mazahua	Latn	585	170	21.3K	8.2K	2.9M	951.7K
tzj	Tz’utujil	Latn	471	136	11.1K	7.3K	1.9M	884.2K
suz	Sunwar	Deva	226	186	226	11.3K	1M	855.2K
knj	W. Kanjobal	Latn	229	126	10.1K	9.2K	1.1M	855K
bim	Bimoba	Latn	410	40	31.1K	6.3K	3.2M	793.4K
gvl	Gulay	Latn	37.9K	126	213K	6.9K	141M	789.2K
bqc	Boko (Benin)	Latn	275	228	9.8K	8.2K	997K	788.4K
tca	Ticuna	Latn	410	117	20K	7.3K	2.3M	786K

pis	Pijin	Latn	1.1K	139	62K	7.2K	7.7M	764K
prk	Parauk	Latn	1.1K	18	12.3K	4.3K	3.1M	734.8K
laj	Lango (Uganda)	Latn	6.5K	144	61K	6.4K	15.8M	730.5K
mel	Central Melanau	Latn	119.3K	103	878.4K	3.7K	315.2M	729.6K
qxr	Cañar H. Quichua	Latn	2.6K	153	40.8K	6.4K	6.6M	724K
niq	Nandi	Latn	26.7K	226	26.7K	4.2K	72.1M	716.2K
ahk	Akha	Latn	244	77	6.2K	4.1K	1.3M	715.5K
shp	Shipibo-Conibo	Latn	874	150	22.4K	3.7K	3.8M	710.4K
hne	Chhattisgarhi	Deva	3K	146	118.4K	4.3K	12M	697K
spp	Supyire Senoufo	Latn	733	123	733	5.8K	4.4M	682.5K
koi	Komi-Permyak	Cyrl	20.7K	196	153.9K	5K	17.1M	664.5K
krj	Kinaray-A	Latn	1.5K	96	54.6K	3.8K	7.6M	616.5K
quf	Lambayeque Quechua	Latn	522	86	8.4K	5.2K	1.5M	609K
luz	S. Luri	Arab	90.5K	354	1.2M	6.7K	329.4M	590.7K
agr	Aguaruna	Latn	465	93	16.1K	3.6K	2.3M	554.5K
tsc	Tswa	Latn	12.6K	82	12.6K	4K	23.4M	521.3K
mqy	Manggarai	Latn	69.3K	119	309K	2.5K	78.9M	506.5K
gof	Gofa	Ethi	2.8K	97	33.8K	5.5K	5.5M	506K
gbm	Garhwali	Deva	2.5K	137	50.8K	3.8K	9.1M	499.6K
miq	Mískito	Latn	236	45	6.4K	3.5K	1.2M	485.6K
dje	Zarma	Latn	913	100	40.2K	3.7K	4.7M	480.7K
awa	Awadhi	Deva	5.8K	126	100.1K	8.4K	11.1M	475K
bjj	Kanauji	Deva	830	107	39.6K	8K	3.1M	439.7K
qvz	N. Pastaza Quichua	Latn	534	88	6.8K	3.5K	1.2M	438.3K
sjp	Surjapuri	Deva	19K	31	498.2K	2.9K	94.3M	430K
tll	Tetela	Latn	200	37	200	2.7K	2.2M	409.8K
raj	Rajasthani	Deva	1.8K	40	1.8K	5.7K	7.1M	405K
kjg	Khmu	Laoo	113	84	3K	2.9K	408.5K	399K
bgz	Banggai	Latn	32K	7	864.1K	17K	79.3M	391.1K
quy	Ayacucho Quechua	Latn	588	78	28.1K	2.7K	4.5M	368.2K
cbk	Chavacano	Latn	10.1K	78	43.8K	2K	10.3M	339.3K
akb	Batak Angkola	Latn	1K	71	21.3K	408	5.2M	337.8K
oj	Ojibwa	Cans	2.5K	135	2.5K	1.6K	9.6M	337.1K
ify	Keley-I Kallahan	Latn	611	79	19.8K	2.8K	2.6M	334K
mey	Hassaniyya	Arab	14.8K	127	109.9K	3K	36.2M	323.5K
ks	Kashmiri	Arab	5.6K	51	53.9K	3.3K	9.4M	320.9K
cac	Chuj	Latn	212	77	3.4K	1.8K	978.7K	319.8K
brx	Bodo (India)	Deva	322	62	5.3K	2.4K	1.1M	304.4K
qup	S. Pastaza Quechua	Latn	169	53	4.3K	2.5K	763.8K	297.8K
syl	Sylheti	Beng	5.9K	61	5.9K	4.3K	21.5M	293.1K
jax	Jambi Malay	Latn	1.5M	58	30M	2.3K	6.8B	290.2K
ff	Fulfulde	Latn	13.6K	26	150K	5K	22.8M	277.6K
ber	Tamazight (Tfng)	Tfng	2.7K	79	12.6K	1.2K	6.4M	265.9K
tks	Takestani	Arab	63.7K	127	63.7K	6.8K	88.9M	260.8K
trp	Kok Borok	Latn	12.8K	36	12.8K	1.7K	29.9M	257.3K
mrw	Maranao	Latn	11.3K	29	11.3K	1K	27.8M	257.2K
adh	Adhola	Latn	2.6K	87	107.2K	1K	14.5M	254.9K
smt	Simte	Latn	1.4K	34	1.4K	703	6.8M	245.4K
srr	Serer	Latn	41.1K	91	41.1K	2.3K	63.6M	240.6K
ffm	Maasina Fulfulde	Latn	1.8K	65	30.1K	2K	4.6M	236.1K

qvc	Cajamarca Quechua	Latn	3.4K	27	14.6K	2.2K	5M	233.7K
mtr	Mewari	Deva	1.8K	11	1.8K	2.2K	7.6M	231.1K
ann	Obolo	Latn	464	56	5K	1.6K	760.9K	215.1K
kaa-Latn	Kara-Kalpak (Latn)	Latn	375.2K	61.2K	3.6M	1.3M	1.5M	209.5K
aa	Afar	Latn	39.5K	32	176.1K	1.3K	63.3M	200K
noe	Nimadi	Deva	2K	22	2K	2.2K	13.8M	195.3K
nut	Nung (Viet Nam)	Latn	29K	67	29K	1.5K	23.5M	184.1K
gyn	Guyanese Creole English	Latn	32.6K	45	211.7K	2.1K	34.5M	177.7K
kwi	Awa-Cuaiquer	Latn	382	37	16.9K	2.2K	1.8M	172.8K
xmm	Manado Malay	Latn	24.5K	58	218.8K	1.2K	48.7M	171.3K
msb	Masbatenyo	Latn	811	41	811	1K	4.4M	167.5K
el-Latn	Greek (Latn)	Latn	-	-	-	-	-	-
doi	Dogri	Deva	-	-	-	-	-	-
mtq	Muong	Latn	-	-	-	-	-	-
dln	Darlong	Latn	-	-	-	-	-	-
cyo	Cuyonon	Latn	-	-	-	-	-	-
abs	Ambonese Malay	Latn	-	-	-	-	-	-
hi-Latn	Hindi (Latn)	Latn	-	-	-	-	-	-
shu	Chadian Arabic	Arab	-	-	-	-	-	-
yaq	Yaqui	Latn	-	-	-	-	-	-
nyo	Nyoro	Latn	-	-	-	-	-	-
cgg	Chiga	Latn	-	-	-	-	-	-
sxu	Upper Saxon	Latn	-	-	-	-	-	-
mdh	Maguindanaon	Latn	-	-	-	-	-	-
rwr	Marwari (India)	Deva	-	-	-	-	-	-
xnr	Kangri	Deva	-	-	-	-	-	-
mui	Musi	Latn	-	-	-	-	-	-
skg	Sakalava Malagasy	Latn	-	-	-	-	-	-
ymm	Maay	Latn	-	-	-	-	-	-
ctd-Latn	Tedim Chin (Latn)	Latn	-	-	-	-	-	-
ayl	Libyan Arabic	Arab	-	-	-	-	-	-
kjb	Q'anjob'al	Latn	-	-	-	-	-	-
rhg-Latn	Rohingya (Latn)	Latn	-	-	-	-	-	-
bmm	N. Betsimisaraka Malagasy	Latn	-	-	-	-	-	-
azg	San Pedro Amuzgos Amuzgo	Latn	-	-	-	-	-	-
kfy	Kumaoni	Deva	-	-	-	-	-	-
bto	Rinconada Bikol	Latn	-	-	-	-	-	-
ja-Latn	Japanese (Latn)	Latn	-	-	-	-	-	-
mfb	Bangka	Latn	-	-	-	-	-	-
ru-Latn	Russian (Latn)	Latn	-	-	-	-	-	-
tuf	Central Tunebo	Latn	-	-	-	-	-	-
ctg	Chittagonian	Beng	-	-	-	-	-	-
pmy	Papuan Malay	Latn	-	-	-	-	-	-
xog	Soga	Latn	-	-	-	-	-	-
te-Latn	Telugu (Latn)	Latn	-	-	-	-	-	-
ber-Latn	Tamazight (Latn)	Latn	-	-	-	-	-	-
mdy	Male (Ethiopia)	Ethi	-	-	-	-	-	-
az-RU	Azerbaijani (Russia)	Cyrl	-	-	-	-	-	-
ta-Latn	Tamil (Latn)	Latn	-	-	-	-	-	-
clu	Caluyanun	Latn	-	-	-	-	-	-

tly-IR	Talysh (Iran)	Arab	-	-	-	-	-	-
ng	Ndonga	Latn	-	-	-	-	-	-
bzc	S. Betsimisaraka Malagasy	Latn	-	-	-	-	-	-
nan-Latn-TW	Min Nan Chinese (Latn)	Latn	-	-	-	-	-	-
ml-Latn	Malayalam (Latn)	Latn	-	-	-	-	-	-
max	North Moluccan Malay	Latn	-	-	-	-	-	-
ar-Latn	Arabic (Latn)	Latn	-	-	-	-	-	-
gom-Latn	Goan Konkani (Latn)	Latn	-	-	-	-	-	-
bg-Latn	Bulgarian (Latn)	Latn	-	-	-	-	-	-
nd	North Ndebele	Latn	-	-	-	-	-	-
zyj	Youjiang Zhuang	Latn	-	-	-	-	-	-
rkt	Rangpuri	Beng	-	-	-	-	-	-
kn-Latn	Kannada (Latn)	Latn	-	-	-	-	-	-
zh-Latn	Chinese (Latn)	Latn	-	-	-	-	-	-
el-CY	Greek (Cypress)	Grek	-	-	-	-	-	-
dcc	Deccan	Arab	-	-	-	-	-	-
bgc	Haryanvi	Deva	-	-	-	-	-	-
mwr	Marwari	Deva	-	-	-	-	-	-
vkt	Tenggarong Kutai Malay	Latn	-	-	-	-	-	-
cr-Latn	Cree (Latn)	Latn	-	-	-	-	-	-
apd-SD	Sudanese Arabic	Arab	-	-	-	-	-	-
trw	Torwali	Arab	-	-	-	-	-	-
bn-Latn	Bengali (Latn)	Latn	-	-	-	-	-	-
gu-Latn	Gujarati (Latn)	Latn	-	-	-	-	-	-
gju	Gujari	Arab	-	-	-	-	-	-
sat-Latn	Santali (Latn)	Latn	-	-	-	-	-	-
ndc-ZW	Ndau	Latn	-	-	-	-	-	-
kmz-Latn	Khorasani Turkish (Latn)	Latn	-	-	-	-	-	-
mr-Latn	Marathi (Latn)	Latn	-	-	-	-	-	-
en-Cyrl	English (Cyrl)	Cyrl	-	-	-	-	-	-
en-Arab	English (Arab)	Arab	-	-	-	-	-	-
ms-Arab	Malay (Jawi)	Arab	-	-	-	-	-	-
ms-Arab-BN	Malay (Jawi, Brunei)	Arab	-	-	-	-	-	-
bhb-Gujr	Bhili	Gujr	-	-	-	-	-	-
pa-Arab	Lahnda Punjabi (PK)	Arab	-	-	-	-	-	-
syl-Latn	Sylheti (Latn)	Latn	-	-	-	-	-	-
ff-Adlm	Fulah	Latn	-	-	-	-	-	-
pcm	Nigerian Pidgin	Latn	-	-	-	-	-	-
tpi	Tok Pisin	Latn	-	-	-	-	-	-
gjk	Kachi Koli	Arab	-	-	-	-	-	-
bfy	Bagheli	Deva	-	-	-	-	-	-
sgj	Surgujia	Deva	-	-	-	-	-	-
nyn	Nyankole	Latn	-	-	-	-	-	-

BCP-47	Name	Script	docs (noisy)	docs (clean)	sents (noisy)	sents (clean)	chars (noisy)	chars (clean)

A.2 Filtering Details

Cursed Substrings Following is the list of cursed substrings that we used to filter the monolingual data. Here are a few general notes about these strings:

1. low quality sentences ending in the pipe character were very common. (Note: this was not Devanagari-script text using a Danda.)

2. The last few regexes are meant to match A N T S P E A K, *List Case*, and weirdly regular text (for instance, lists of shipping labels or country codes)

Here is the complete list of cursed substrings and cursed regexes, along with the function used for filtering:

```
# this implementation is for demonstration and not very efficient;
# to speed it up, use string inclusion (`in`) instead of regex for
# all but the last four, and for those use a compiled regex.
def is_cursed(s):
  return any(re.findall(curse, s) in s for curse in CURSED_SUBSTRINGS)


CURSED_SUBSTRINGS = [' \u2116', '\ufffd\ufffd\ufffd', '\\|\\s*$',
' nr\\.$',  'aute irure dolor ', ' sunt in culpa qui ',
'orem ipsum ', ' quis nostrud ', ' adipisicing ', '
dolore eu ', ' cupidatat ', 'autem vel eum',
'wisi enim ad', ' sex ', ' porn ',
'\u9ec4\u8272\u7535\u5f71', 'mp3', 'ownload',
'Vol\\.', ' Ep\\.', 'Episode', ' \u0433\\.\\s*$', ' \u043a\u0433\\.\\s*$',
' crusher ', ' xxx ',
' ... ... ... ... ... ... ... ... ...',
```

```
' .... .... .... .... .... .... .... .... ....',
' [^ ] [^ ] [^ ] [^ ] [^ ] [^ ] [^ ] [^ ] [^ ]',
', ..,,? ..,,? ..,,? ..,,?',
]
```

Virama Correction Below is the virama substitution code:

```
_VIRAMA_CHARS = (
'\u094d\u09cd\u0a4d\u0acd\u0b4d\u0bcd\u0c4d\u0ccd\u0d3b'
'\u0d3c\u0d4d\u0dca\u0e3a\u0eba\u0f84\u1039\u103a\u1714'
'\u1734\u17d2\u1a60\u1b44\u1baa\u1bab\u1bf2\u1bf3\u2d7f'
'\ua806\ua82c\ua8c4\ua953\ua9c0\uaaf6\uabed\u10a3f\u11046'
'\u1107f\u110b9\u11133\u11134\u111c0\u11235\u112ea\u1134d'
'\u11442\u114c2\u115bf\u1163f\u116b6\u1172b\u11839\u1193d'
'\u1193e\u119e0\u11a34\u11a47\u11a99\u11c3f\u11d44\u11d45'
'\u11d97\u1031\u1057\u1058\u1059\u1056\u1060\u1062\u1068'
'\u1063\u1067\u1068\u1069\u105e\u105f\u1036\u103d\u102d'
'\u102f\u102e\u102d\u1030\u1033\u1034\u1035\u102f\u1032'
'\u102c\u103c\u103d\u103e\u102b\u1037\u1038\u25cc\u25cc'
'\u000a\u1071\u1072\u1073\u1074\u1082\u1083\u1084\u1085'
'\u1086\u1087\u1088\u1089\u108a\u108b\u108c\u108d\u108f'
'\u109a\u109b\u109c\u109d\ua9e5\uaa7b\uaa7c\uaa7d'
)


def remove_viramas(x: str) -> str:
  return '%s' % regex.sub(r' ([%s]) ' % _VIRAMA_CHARS, '\\1', x)
```

Chinese Porn Filter Below is the Chinese porn filter list:

```
zh_pornsignals = [
'caoporn', 'caoprom', 'caopron', 'caoporen', 'caoponrn',
'caoponav', 'caopom',
'caoorn', '99re', 'dy888', 'caopro', 'hezyo', 're99',
'4438x', 'zooskool',
'xfplay', '7tav', 'xxoo', 'xoxo', '52av', 'freexx',
'91chinese', 'anquye',
'cao97', '538porm', '87fuli', '91pron', '91porn',
'26uuu', '4438x', '182tv',
'kk4444', '777me', 'ae86', '91av', '720lu', 'yy6080',
'6080yy', 'qqchub',
'paa97', 'aiai777', 'yy4480', 'videossexo', '91free',
'\u4e00\u7ea7\u7279\u9ec4\u5927\u7247',
'\u5077\u62cd\u4e45\u4e45\u56fd\u4ea7\u89c6\u9891',
'\u65e5\u672c\u6bdb\u7247\u514d\u8d39\u89c6\u9891\u89c2\u770b',
'\u4e45\u4e45\u514d\u8d39\u70ed\u5728\u7ebf\u7cbe\u54c1',
'\u9ad8\u6e05\u6bdb\u7247\u5728\u7ebf\u770b',
'\u65e5\u672c\u6bdb\u7247\u9ad8\u6e05\u514d\u8d39\u89c6\u9891',
'\u4e00\u7ea7\u9ec4\u8272\u5f55\u50cf\u5f71\u7247',
'\u4e9a\u6d32\u7537\u4eba\u5929\u5802',
'\u4e45\u4e45\u7cbe\u54c1\u89c6\u9891\u5728\u7ebf\u770b',
'\u81ea\u62cd\u533a\u5077\u62cd\u4e9a\u6d32\u89c6\u9891',
'\u4e9a\u6d32\u4eba\u6210\u89c6\u9891\u5728\u7ebf\u64ad\u653e',
'\u8272\u59d1\u5a18\u7efc\u5408\u7ad9',
'\u4e01\u9999\u4e94\u6708\u556a\u556a',
'\u5728\u7ebf\u89c6\u9891\u6210\u4eba\u793e\u533a',
```

'\u4e9a\u6d32\u4eba\u6210\u89c6\u9891\u5728\u7ebf\u64ad\u653e',
'\u4e45\u4e45\u56fd\u4ea7\u81ea\u5077\u62cd',
'\u4e00\u672c\u9053',
'\u5927\u9999\u8549\u65e0\u7801',
'\u9999\u6e2f\u7ecf\u5178\u4e09\u7ea7',
'\u4e9a\u6d32\u6210\u5728\u4eba\u7ebf\u514d\u8d39\u89c6\u9891',
'\u5929\u5929\u8272\u7efc\u5408\u7f51',
'\u5927\u9999\u8549\u4f0a\u4eba\u4e45\u8349',
'\u6b27\u7f8e\u4e00\u7ea7\u9ad8\u6e05\u7247',
'\u5929\u5929\u9c81\u591c\u591c\u556a\u89c6\u9891\u5728\u7ebf',
'\u514d\u8d39\u9ec4\u7247\u89c6\u9891\u5728\u7ebf\u89c2\u770b',
'\u52a0\u6bd4\u52d2\u4e45\u4e45\u7efc\u5408',
'\u4e45\u8349\u70ed\u4e45\u8349\u5728\u7ebf\u89c6\u9891',
'\u97e9\u56fd\u4e09\u7ea7\u7247\u5927\u5168\u5728\u7ebf\u89c2\u770b',
'\u9752\u9752\u8349\u5728\u7ebf\u89c6\u9891',
'\u7f8e\u56fd\u4e00\u7ea7\u6bdb\u7247',
'\u4e45\u8349\u5728\u7ebf\u798f\u5229\u8d44\u6e90',
'\u556a\u556a\u556a\u89c6\u9891\u5728\u7ebf\u89c2\u770b\u514d\u8d39',
'\u6210\u4eba\u798f\u5229\u89c6\u9891\u5728\u7ebf\u89c2\u770b',
'\u5a77\u5a77\u6211\u53bb\u4e5f',
'\u8001\u53f8\u673a\u5728\u7ebf\u56fd\u4ea7',
'\u4e45\u4e45\u6210\u4eba\u89c6\u9891',
'\u624b\u673a\u770b\u7247\u798f\u5229\u6c38\u4e45\u56fd\u4ea7',
'\u9ad8\u6e05\u56fd\u4ea7\u5077\u62cd\u5728\u7ebf',
'\u5927\u9999\u8549\u5728\u7ebf\u5f71\u9662',
'\u65e5\u672c\u9ad8\u6e05\u514d\u8d39\u4e00\u672c\u89c6\u9891',
'\u7537\u4eba\u7684\u5929\u5802\u4e1c\u4eac\u70ed',
'\u5f71\u97f3\u5148\u950b\u7537\u4eba\u8d44\u6e90',

'\u4e94\u6708\u5a77\u5a77\u5f00\u5fc3\u4e2d\u6587\u5b57\u5e55',
'\u4e9a\u6d32\u9999\u8549\u89c6\u9891\u5728\u7ebf\u64ad\u653e',
'\u5929\u5929\u556a\u4e45\u4e45\u7231\u89c6\u9891\u7cbe\u54c1',
'\u8d85\u78b0\u4e45\u4e45\u4eba\u4eba\u6478\u4eba\u4eba\u641e',
]

A.3 Other issues fixed after the self-audit

Consulting Language Speakers For a few languages, we had strong suspicions that the text was noisy or spurious, but were unable to acertain the quality of the data. In these cases we asked a native speaker to audit the data. Based on their recommendations, we did the following:

1. `zh`, `zh_Latn`: This resulted in the special filters described below.
2. `en_Arab`, `tly_IR`: This data was found to boilerplate, so we removed this data.
3. `fa`, `bho`: No changes were made.

Language Renames and Merges For several languages, we found that (mostly by checking URLs) the corpora were in languages different from the LangID predictions. This led to the following changes:

1. dty renamed to `zxx-xx-dtynoise`, aka a "language" of noise. This is mainly mis-rendered PDFs and may have practical applications for denoising, or for decoding such garbled PDFs.
2. `fan` renamed to `bum`
3. `cjk` merged into the `gil` dataset
4. `bjj` merged into the `awa` dataset
5. `ss-SZ` renamed to `ss` – this was a result of inconsistent data labels.

A.4 Monolingual Data Details

Notes from rounds 2 and 3 of the self-audit can be seen in Table A.2. Some of these notes may refer to previous, less filtered versions of the data, especially those with a "r1" (meaning "round 1").

Some of them however do have some useful information about quirks of problems with the current dataset. The overall statistics of MADLAD-400 are in Table A.1.

Table A.2: Notes that we made about individual samples while auditing them. Some languages have notes from the earlier round of auditing in parentheses, e.g. '(r1: get this checked by Hindi speaker)'. Notes from Round 0, which were used to find cursed substrings, were not kept.

BCP-47	notes
aa	some pretty bad data but also some good data. filter on "Woo" (case sensitive) (r1: ok)
abs	all short nonsense remove (r1: ok)
abt	fine; bible (r1: ok)
ace	good; bible (r1: ok)
acf	good; bible (r1: ok)
ach	good; bible (r1: ok)
ada	good; bible; likely mixed with gaa (r1: ok but odd character usage LATIN CAPITAL LETTER OPEN O when it should be lower case in the middle of words)
adh	good; bible (r1: ok, lots bible)
ady	good (r1: ok but weird boilerplate)
af	good (r1: ok)
agr	good; bible (r1: ok; some AL in Arabic script)
ahk	good; bible; crazy diacritics (r1: ok but weird: lots of u748 MODIFIER LETTER VOICING)
ak	good; much but not all bible (r1: ok)
akb	good; bible (r1: empty)
alt	WAIT THIS IS AMAZING IT IS ACTUALLY ALTAI! e.g. from urls like https://altaicholmon.ru/2020/02/28/jarashty-la-jajaltany-jarkyndu-lekeri/ (r1: ok but there are just lots of numbers...not very clean)
alz	good; bible (r1: ok; bible)
am	good (r1: ok)
amu	good; bible; crazy diacritics (r1: empty)
ang	much noise but some good Old English in there! (r1: ok; wikipedia; one document that is just "lastfootwear.com" 1M times)
ann	good; all from wikimedia incubator (r1: ok)
apd-SD	terribly questionable; probably remove (r1: maybe ok, but looks like lots of template....maybe remove)
ape	good; bible (r1: remove)
ar	good (r1: ok)
ar-Latn	terrible, 0pct correct, remove (r1: remove)
arn	good; bible (r1: ok)
as	good (r1: ok)
av	good (r1: ok)
awa	OK; should be used with caution and suspicion (r1: remove)
awa	all bible in awadhi (awa). Renamed from bjj (r1: remove)
ay	good; mix of bible and other news sources (r1: ok but very noisy)
ayl	remove. not ayl. (r1: uh this is all Arabic with "homo" in English...remove?)
az	good (r1: ok)
az-RU	good; a lot of JW (r1: ok)
azg	70pct short noise; 30pct good bible (r1: empty)

ba	ok (r1: ok)
ban	ok bible (r1: ok)
bas	ok; has some fun blog stuff! (r1: empty)
bbc	ok (r1: ok)
bci	ok bible (r1: ok; bible)
be	ok (r1: ok)
ber	ok great! (r1: ok; Mixed in French)
ber-Latn	ok (r1: ok)
bew	mostly blogs. i have no way of knowing if this is standard indonesian or not (r1: ok; noisy)
bfy	very bad. remove unless it looks better after filtering short docs; remove (r1: remove)
bg	ok (r1: ok)
bg-Latn	ok (r1: ok but questionable...Slavic speaker review needed)
bgc	super sketch. Remove unless short doc filter leaves some. remove (r1: very questionable....Hindi speaker review)
bgp	almost all ur-Latn. consider removing or renaming (r1: very questionable. Remove? mainly ur-Latn)
bgz	idk maybe ok but probably bad (r1: remove. Wow, this is amazing. It is in all sorts of languages – the only thing they share is that they each have like 500 question marks)
bhb-Gujr	bad. remove. all junk gu. (r1: remove; great noise?)
bho	mostly from anjoria.com. Ankur reviews and says that it looks like valid Bhojpuri for the most part (r1: questionable but ok?)
bi	good! fun! (r1: ok)
bik	ok. keep in mind the bik vs bcl issue. (r1: ok)
bim	good; bible (r1: empty)
bm	good (r1: ok but these headers are LONG)
bmm	terrible. filter on short and reevaluate (r1: remove)
bn	ok (r1: ok)
bn-Latn	ok (r1: ok)
bo	needs some serious script filtering. but there is some ok data in there. (r1: ok)
bqc	ok; bible (r1: ok but too short?)
br	ok after shortfilter (r1: ok)
bru	ok; bible (r1: ok)
brx	quite good! (r1: ok but questionable...Hindi speaker review?)
bs	good (r1: ok)
bto	bad; remove unless short filter keeps enough (r1: empty)
bts	ok; mostly bible (r1: ok)
btx	ok probably (r1: ok)
bua	ok (r1: ok)
bum	ok bible; but technically wrong language. Data is in Bulu, not Fang, though they are closely related, so ranamed from "fan"
bus	ok; bible; about 50bzc (r1: ok bible)
bzj	ok bible (r1: ok)
ca	ok (r1: ok i guess....but is it actually italian?)
cab	ok jw (r1: ok)
cac	ok bible (r1: ok bible)
cak	ok bible (r1: ok bible)
cbk	ok bible; not Spanish (r1: remove; all Spanish)
cce	ok jw (r1: empty)
ce	ok (r1: ok)
ceb	ok (r1: ok)
cfm	ok mostly from chinland.co (r1: ok)
cgg	rather noisy but potentialy ok. not sure if WL or not (r1: ok)
ch	ok; not sure about WL (r1: ok)
chk	ok bible (r1: ok bible)

chm	ok; fyi watch out for yandex translationese (r1: ok)
chr	ok bible (r1: ok)
ckb	ok (r1: ok)
clu	ok bible (r1: ok)
cnh	good, some local news! not sure if WL (r1: ok)
cni	ok; bible; lots of mixed in content in not,cob,cpc,arl (r1: ok)
co	ok;l i suspect lots of MT (r1: ok i guess?)
cr-Latn	noise and lorem ipsom. But some ok Cree text. (r1: mostly Lorem Ipsom. remove? Or release with note? there is some plausible stuff here too.)
crh	ok (r1: ok but review with russian speaker as it could be russian....)
crs	ok (r1: ok)
cs	ok (r1: ok)
ctd-Latn	ok; from some local news? (r1: ok)
ctg	probably terrible probably remove (r1: very questionable....remove?)
ctu	ok bible (r1: ok bible)
cuk	ok bible (r1: ok bible)
cv	good (r1: ok)
cy	ok after shortfilter; OK (r1: ok)
cyo	terrifying noise; remove (r1: empty)
da	ok (r1: ok)
dcc	remove (r1: empty)
de	ok (r1: ok)
din	ok after short doc filter (r1: ok but LONG headers uh oh)
dje	ok; mostly but not all bible (r1: ok bible)
djk	ok; bible+jw (r1: empty)
dln	ok bible (r1: empty)
doi	ok actually nice! (r1: sus; review by hindi speaker)
dov	ok bible + jw (r1: ok)
dtp	ok; mostly from www.newsabahtimes.com.my (r1: ok)
dv	good (r1: ok)
dwr	ok; bible; mixed script (r1: empty)
dyu	ok bible (r1: empty)
dz	ok; hidden parallel text; maybe actually bo; mainly buddhist (r1: ok; mixed dz-Latn)
ee	good; mostly religious (r1: ok bible)
el	ok (r1: ok)
el-CY	bad (r1: v suspicious; mainly comma lists or boilerplate; remove)
el-Latn	good; a lot of old content! (r1: ok)
emp	ok bible (r1: ok)
en	ok (r1: ok)
en-Arab	Ali reviewed; this is not good data. remove. (r1: idk review w/Arabic reader)
en-Cyrl	ok ... some fr-Cyrl too and maybe others (r1: OMG LOL yes ok)
enq	ok bible (r1: ok bible)
eo	ok; likely a lot of MT (r1: ok)
es	good (r1: ok)
et	ok (r1: ok)
eu	ok (r1: ok; lots of poetry?)
fa	consulted Ali; he says it's ok (r1: ok)
ff	ok after shortfilter (r1: some noise but some nice stuff! ok!)
ff-Adlm	good (r1: ok sweet)
ffm	ok bible; mixed fulfulde dialects; consider mergind with ff (r1: ok but idk the dialect)

fi	ok (r1: ok but lotsa headers)
fil	ok more bible than expected for such a major language (r1: ok pls note in release that this is the same as tl)
fip	ok jw ; but wrong language. mostly Mambwe-Lungu and Bemba, not Fipu (mgr+bem vs fip) (r1: ok bible)
fj	ok (r1: ok bible lotsa noise)
fo	good (r1: ok TODO check that this is not icelandic review)
fon	ok mostly jw but not all (r1: ok bible)
fr	ok (r1: ok)
frp	fair amount from wikipedia. (r1: remove; all noise + hashtags)
fy	ok plausible but i bet there is a lot of Dutch in there (r1: ok)
ga	ok some en noise (r1: ok)
gag	has 1-2 cyrillic examples with small amts of arabic script noise (r1: ok)
gbm	ok (r1: ok)
gd	ok (r1: ok; but barely)
gil	empty; but merged in data in "cjk" (r1: empty)
gil	this is all in gil (Kiribati). merged into "gil" (r1: empty)
gjk	empty remove (r1: empty)
gju	remove short boilerplate (r1: empty)
gl	ok (r1: ok)
gn	ok some broken characters some bible (r1: ok)
gof	ok some bible (r1: empty)
gom	ok (r1: ok)
gom-Latn	filter on really short boilerplate in en; some porn; after: ok very noisy ; some ok stuff ; release with disclaimer (r1: ok)
gor	ok bible (r1: ok)
grc	warning: this is likely polyphonic greek, not ancient greek (r1: ok but idk diff between ancient and modern greek)
gsw	wtf is happening here; keep with disclaimer; STILL BOILERPLATE (r1: ok but idk diff between gsw and de)
gu	ok (r1: ok some en boilerplate)
gu-Latn	filter short en boillerplate and repetitive sentences (r1: lots of social media pages and some porn)
gub	ok bible (r1: empty)
guc	ok bible (r1: ok)
guh	ok bible (r1: ok)
gui	ok bible (r1: ok)
gv	filter short repetitive sentenecs; still same but keep (r1: ok)
gvl	filter short boilerplate mostly bible (r1: ok)
gym	ok biblle (r1: ok)
gyn	remove boilerplate and porn (r1: remove)
ha	ok (r1: ok)
haw	ok scam tv products (r1: ok but filter u65533 REPLACEMENT CHARACTER)
hi	ok some porn (r1: ok but some en boilerplate)
hi-Latn	filter porn this is half porn (r1: ok but some hi and en)
hif	ok some en noise and religious (r1: ok it is in Latin)
hil	ok some en boilerplate (r1: ok)
hmn	ok (r1: ok)
hne	ok (r1: ok)
ho	ok (r1: ok but but split between wiki boilerplate and actual content)
hr	ok (r1: ok)
ht	ok (r1: ok)
hu	ok (r1: ok)
hui	ok some bible (r1: ok bible)
hus	ok bible (r1: some wiki boilerplate)
hvn	ok religioous text (r1: ok bible)

hy	ok (r1: ok)
iba	ok jw data (r1: ok)
ibb	ok bible and repeated @ (r1: ok but bible and some repeated lines)
id	ok (r1: ok)
ify	ok bible (r1: empty)
ig	ok (r1: ok)
ilo	ok some bible (r1: some repetitive content)
inb	ok bible (r1: remove; it's a single bible doc lol)
is	ok (r1: ok)
iso	ok jw (r1: ok)
it	ok (r1: ok)
iu	filter script some is en rest is iu script (r1: ok filter latin script)
ium	filter out zh (r1: remove mostly en)
iw	ok (r1: ok has some codemixing because of boilerplate)
izz	ok bible (r1: empty)
ja	ok a little en mixed in (r1: ok but some porn)
ja-Latn	remove maybe low quality short and repeated (r1: ok some noise that is manga pages in english)
jac	ok bible (r1: remove 'home loan' repeated over and over)
jam	ok bible (r1: ok)
jax	filter mostly text.medjugorje.ws boilerplate (r1: remove)
jiv	ok bible
jv	ok (r1: ok)
jvn	ok bible (r1: ok)
ka	ok (r1: ok)
kaa	ok (FYI cyrllic) (r1: ok)
kaa-Latn	ok urls are .ru or .kz (r1: ok)
kac	ok (r1: ok)
kbd	ok many .ru (r1: ok some repetitive text and en noise)
kbp	not sure if right script wiki says latin (r1: ok)
kek	ok jw bible (r1: ok bible)
kfy	filter virama issue (r1: ok)
kg	ok bible jw (r1: ok)
kha	ok (r1: ok some repetitive boilerplate)
kj	ok (r1: filter english out)
kjb	ok bible (r1: empty)
kjg	ok bible (r1: empty)
kjh	ok .ru domain (r1: ok)
kk	ok (r1: ok)
kl	ok (r1: ok)
km	ook (r1: ok)
kmb	ok bible jw (r1: ok)
kmz-Latn	ok soome ar script noise (r1: ok)
kn	ok (r1: ok)
kn-Latn	filter en noise of karnataka govt websites (r1: filter porn there is too much porn and repetitive content)
knj	ok bible (r1: empty)
ko	ok (r1: ok)
koi	ok (r1: ok)
kos	ok lds bible (r1: ok bible)
krc	ok (r1: ok some repetitive content)
kri	ok boilerplate noise bible jw (r1: remove repetitive)

ks	ok shorter docs (r1: ok)
ksd	ok bible (r1: ok bible)
ksw	ok bible (r1: ok)
ktu	ok bible jw (r1: ok)
ku	ok (r1: ok)
kum	ok (r1: ok)
kv	ok a lil boilerplate vibes (r1: ok)
kw	ok short boilerplate bible wiki; ok some porn (r1: ok filter english)
kwi	ok bible (r1: ok)
ky	ok (r1: ok)
la	ok some broken chars
laj	ok bible
lb	ok shorter text; ok AFTER
lg	ok lot of www.bukedde.co.ug in this
lhu	ok bible
ln	ok bible jw
lo	ok many entities in latin script
lrc	ok
lt	ok
ltg	ok mostly www.lakuga.lv
lu	ok jw
lus	ok
luz	terrible; remove
lv	ok
mad	remove mostly short text
mag	ok fix virama issue
mai	ok mild amounts of en noise
mak	ok bible
mam	ok bible jw
mas	ok some amount of bible
max	remove short some ru
maz	ok bible jw
mbt	ok bible
mdf	ok some short docs
mdh	filter porn short text and repetitive boilerplate
mdy	ok bible
mel	remove noisy en
meo	ok mostly blogs
meu	ok bible
mey	mostly short and noisy borderline
mfb	remove short boilerplate
mfe	ok mostly bible maybe some french creole short doc noise
mg	ok some bible jw
mgh	ok bible jw
mh	ok jw lds
mi	ok
min	ok mostly wiki and bible
miq	ok
mk	ok
mkn	ok bible

ml	ok
ml-Latn	ok some short docs
mn	ok
mni	ok
mnw	remove en noise and boilerplate
mps	ok bible
mqy	bible remove short docs
mr	ok fix virama
mr-Latn	remove mostly porn and short docs
mrj	remove short docs; ok
mrw	ok remove short docs
ms	ok
ms-Arab	ok mostly utusanmelayu website
ms-Arab-BN	ok not sure if same as ms-Arab
msb	ok bible
msi	ok filter short docs
msm	ok bible
mt	ok
mtq	remove short doc repetitive
mtr	ok fix virama remove en noise
mui	remove short docs
mwr	filter short docs fix virama
my	filter noise and en fix virama
myv	maybe has .ru urls
nan-Latn-TW	ok
nd	ok
ndc-ZW	ok
ne	ok
new	ok
ng	ok
ngu	ok
nhe	ok
nia	ok
nij	ok
niq	ok
nl	ok
nnb	ok
no	ok
noa	ok
noe	ok
nog	ok
nr	ok
nso	ok
nut	ok
nv	ok
ny	ok
nyn	ok
nyo	ok
nyu	ok
nzi	ok

oc	ok
oj	ok
om	ok
or	ok
os	ok
otq	ok
pa	ok
pa-Arab	ok
pag	bible
pam	remove
pap	ok
pau	ok
pck	ok
pcm	ok
pis	bible
pl	ok
pmy	remove
pon	bible
ppk	bible
prk	ok
ps	ok
pt	ok
qu	ok
qub	bible
quc	bible
quf	bible
quh	bible
qup	bible
quy	bible
qvc	bible
qvi	bible
qvz	bible
qxr	bible
raj	ok
rcf	ok
rhg-Latn	remove
rki	ok
rkt	ok
rm	ok
rmc	ok
rn	bible
ro	ok
rom	bible
ru	ok
ru-Latn	ok
rw	ok
rwo	bible
rwr	remove
sa	ok
sah	ok

sat-Latn	good! all from local news sources
sd	good
sda	ok bible
se	good
seh	ok jw
sg	ok jw
sgj	remove
shn	mostly English boilerplate. filter by latin text before releasing
shp	ok bible
shu	quite questionable. prob remove
si	good
sja	ok bibe
sjp	terible; probably remove; check again after short filter
sk	ok
skg	terrible; remove
skr	ok; some pnb mixed in
sl	ok
sm	ok
smt	ok bible but lots of different bibles!
sn	ok
so	good
spp	ok bible
sq	good
sr	ok
srm	ok; bible + jw
srn	ok bible + jw
srr	remove; englishboilerplate
ss	good mix of data ; renamed from "ss"
st	ok
stq	ok i think ?
su	good
sus	hella sus jk ok bible
suz	ok bible
sv	ok
sw	ok
sxn	ok bible ; also wild diacritics
sxu	rvisit after shortfilter
syl	idk maybe ok ?
syl-Latn	revist or remove after shortfilter
syr	good; practictitioners should keep dialect in mind.
ta	ok
ta-Latn	good text .... but pornographic, like all Indic-Latn datasets
tab	idk plausibly ok
taj	ok bible
tbz	good mostly bible but not all
tca	ok bible + jw
tcy	good; mostly wikipedia; likely some konkani mixed in
tdx	ok jw
te	ok a lot of weirdly low quality looking content like commerce
te-Latn	great good text....but all pornographic stories + blogs, like all Indic-Latn text

teo	ok bible
tet	good ; actually a lot of fun data!
tg	good
th	ok
ti	ok; poor tigray
tiv	ok jw
tk	ok; a few weird docs
tks	ok bible but again i think some mixed dialects
tlh	ok, but why tf are there websites in klingon? all MT ?
tll	ok jw
tly-IR	deeply sus; remove after shortfilter
tn	good
to	good ; news bible government
toj	ok jw
tpi	empy
tr	ok
trp	good ; lots of random stuff
trw	sus; remove
ts	good
tsc	ok
tsg	much noise but some good data too!
tt	good plus some nonunicode misrendered PDF
tuc	ok bible
tuf	ok bible
tvl	ok jw
twu	ok bible, but also i think it’s lots of mixed similar dialects
tyv	ok fun stuff plus some russian noise i think
tyz	ok bible bu again i think some mixed dialects
tzh	ok jw
tzj	ok bible
tzo	ok bible + jw
ubu	ok bible
udm	ok
ug	ok
uk	ok
ur	ok
uz	ok some cyrllic noise
ve	ok mostly bible jw
vec	very noisy has wiki from other langs and .it websites so not sure if vec
vi	ok
vkt	1 doc remove
wa	ok lots of wiki stuff
wal	ok bible + jw
war	ok but v sus. Pls filter out wikipedia
wo	ok; mostly bible. PS i have found that Wolof web-crawled data is often bad, so pls give an extra look if you want
xal	ok has .ru sites though
xh	ok
xmm	very noisy lots of dj tiktok and peppa pig repeated
xnr	ok maybe fix virama though it seems fine
xog	ok bible and stories

yap	ok
yaq	remove
yi	ok
ymm	remove
yo	ok
yua	ok
yue	pretty low quality; mostly not Canto
za	revisit after shortfilter
zap	ok JW. PS pls note that at least some Zapotec speakers view it as one language, not as a million dialects like ISO does
zh	mixed simplified and trad; also much porn
zh-Latn	revisit after shortfilter
zne	ok jw
zu	good
zxx-xx-dtynoise	BEAUTIFUL NOISE rename but keep as beautiful xample. (was called "dty")
zyj	deeply bad data .. revisit after shortfilter
zza	good ; also pls note that the Zazaki community is often super into NLP for Zazaki

A.5 Parallel Data Details

To create the dataset described in Section 2.1.3, we use the data sources described in Table A.3. After preprocessing, we obtain a dataset with a total of 157 different languages and 4.1B sentence pairs that vary from `en-es` with 280.3M sentence pairs to `zu-en` with 7959 sentence pairs. The list of language pairs along with the associated data count is available along with the model checkpoints.

Table A.4: Preprocessed multiway data divided by target language in decreasing order of number of # Sentence Pairs.

Language (Code)	# Sentence Pairs	Language (Code)	# Sentence Pairs
English (en)	6825073150	Malagasy (mg)	1190715
Spanish (es)	630962081	Punjabi (pa)	1122414
German (de)	542301841	Hausa (ha)	952163
French (fr)	540506439	Latin (la)	882038
Portuguese (pt)	355008897	Inuktitut (iu)	841459
Italian (it)	283297447	Myanmar (Burmese) (my)	759196
Dutch (nl)	272264696	Walloon (wa)	706400
Swedish (sv)	183450145	Uzbek (uz)	695915
Czech (cs)	166419080	Luxembourgish (lb)	645610
Polish (pl)	154989834	Assamese (as)	623321
Mandarin Chinese (zh)	147268699	Pashto (ps)	606852
Danish (da)	145764833	Armenian (hy)	603397

Continued on next page

Table A.4 continued from previous page

Language (Code)	# Sentence Pairs	Language (Code)	# Sentence Pairs
Hungarian (hu)	136980377	Sindhi (sd)	597211
Russian (ru)	134568973	Northern Sami (se)	585304
Romanian (ro)	117227121	Bashkir (ba)	551617
Slovak (sk)	102889505	Amharic (am)	548828
Finnish (fi)	100468712	Somali (so)	543643
Greek (el)	97855862	Dhivehi (dv)	543145
Arabic (ar)	83364997	Kurdish (Kurmanji) (ku)	483477
Bulgarian (bg)	68424658	French (Canada) (fr_CA)	472160
Lithuanian (lt)	66562681	Odia (Oriya) (or)	465494
Slovenian (sl)	65876962	Faroese (fo)	434288
Latvian (lv)	59061871	Kannada (kn)	417255
Norwegian (no)	57783339	Kinyarwanda (rw)	386133
Estonian (et)	47403866	Wu Chinese (wuu)	369891
Japanese (ja)	42480849	Lombard (lmo)	366147
Korean (ko)	33987656	Egyptian Arabic (arz)	353218
Catalan (ca)	28627098	Uyghur (ug)	337582
Croatian (hr)	28475191	Limburgan (li)	304279
Turkish (tr)	26508188	Aragonese (an)	251004
Ukrainian (uk)	26258530	Sicilian (scn)	249682
Icelandic (is)	24073502	Dzongkha (dz)	248172
Persian (fa)	23237370	Meiteilon (Manipuri) (mni)	242426
Indonesian (id)	19860783	Maori (mi)	234833
Vietnamese (vi)	18803243	Nuer (nus)	231194
Hebrew (he)	17171755	Magahi (mag)	230839
Macedonian (mk)	15900226	Bhojpuri (bho)	230378
Irish (ga)	14939080	Shan (shn)	229967
Galician (gl)	14305858	Friulian (fur)	228928
Serbian (sr)	13989182	Kashmiri (ks)	228580
Albanian (sq)	13904626	Sardinian (sc)	227585
Basque (eu)	13755514	Kanuri (kr_Arab)	227549
Maltese (mt)	12583626	Dinka (din)	227233
Esperanto (eo)	11431440	Venetian (vec)	225065
Hindi (hi)	11044239	Chhattisgarhi (hne)	224085
Bosnian (bs)	10906114	Ligurian (lij)	222513
Serbo-Croatian (sh)	10592606	Central Atlas Tamazight (tzm)	222365
Bengali (bn)	8394639	Tamasheq (taq_Tfng)	220907
Sinhala (si)	7363378	Dari (prs)	220899
Thai (th)	6623168	Banjar (bjn)	220753
Malayalam (ml)	5995504	Achinese (ace_Arab)	220651
Tamil (ta)	5552495	Tamasheq (taq)	219708
Marathi (mr)	5517596	Banjar (bjn_Arab)	219020
Telugu (te)	5379834	Achinese (ace)	217062
Malay (ms)	5241831	Bambara (bm)	216698
Filipino (fil)	4532364	Balinese (ban)	215965
Urdu (ur)	4526509	Moroccan Arabic (ary)	214226
Taiwanese Mandarin (zh_Hant)	3782804	Nigerian Fulfulde (fuv)	209659

Continued on next page

Table A.4 continued from previous page

Language (Code)	# Sentence Pairs	Language (Code)	# Sentence Pairs
Swahili (sw)	3220515	Silesian (szl)	209224
Nepali (ne)	2833740	Kashmiri (ks_Deva)	208575
Norwegian Nynorsk (nn)	2726466	Buginese (bug)	207209
Azerbaijani (az)	2353731	Guarani (gn)	201241
Kazakh (kk)	2153356	Latgalian (ltg)	200302
Tajik (tg)	2005810	Crimean Tatar (crh_Latn)	192342
Low German (nds)	1942746	Kanuri (kr)	191037
Occitan (oc)	1934958	Scottish Gaelic (gd)	162882
Georgian (ka)	1882623	Bavarian (bar)	159419
Belarusian (be)	1808663	Javanese (jv)	144974
Tatar (tt)	1763087	Mongolian (mn)	132599
Western Frisian (fy)	1500410	Zulu (zu)	130485
Afrikaans (af)	1459139	Kyrghyz (ky)	128832
Breton (br)	1431191	Low German (Netherlands) (ndsNL)	113101
English (simple) (en_xx_simple)	1413074	Mirandese (mwl)	96829
Khmer (km)	1330590	Ido (io)	72811
Gujarati (gu)	1309569	Igbo (ig)	41296
Cebuano (ceb)	1255728	Turkmen (tk)	40650
Welsh (cy)	1229401	Yiddish (yi)	37128
Xhosa (xh)	1219701	Yoruba (yo)	26589
		Total # Sentence Pairs	11944961985

A.6 Language Codes

The specifics of the language code changes described in Section 2.1.4.1 that we made are as follows:

1. We use fil for Filipino/Tagalog, not tl

2. We use ak for Twi/Akan, rather than tw. This includes Fante.

3. Unfortunately, we use the macro code chm for Meadow Mari (instead of the correct mhr), and mrj for Hill Mari

4. By convention, we use no for Norwegian Bokmål, whereas some resources use nb

5. By convention we use ps for Pashto instead of pbt (Southern Pashto)

6. By convention, we use ms for Standard Malay, not zlm

Table A.3: The various data sources used to create the parallel data used to train our MT models with the number of available languages and language pairs. (*for NewsCommentary v14 we only use Kazakh (`kk`) data)

Dataset	Version # Language Pairs	# Languages	
Europarl Koehn (2005)	v7	40	21
	v9	12	8
Paracrawl Esplà-Gomis et al. (2019)	-	64	41
TED57 Ye et al. (2018)	-	114	58
Tanzil Tiedemann (2012)	-	1560	40
NewsCommentary sta	v10	6	3
	v14*	2	1
	v16	200	15
Wikimatrix Schwenk et al. (2019)	-	1620	85
Wikititles sta	v2	12	14
OPUS100 Zhang et al. (2020)	-	198	100
SETimes Tiedemann (2012)	-	90	10
UNv1.0 Ziemski et al. (2016)	-	20	5
Autshumato Groenewald & Fourie (2009)	-	10	5
PMIndia Haddow & Kirefu (2020)	v1	13	13
CVIT Philip et al. (2019)	MKB (v0), PIB (v1.3)	110	11
Inuktitut Joanis et al. (2020)	-	2	1
NLPC Fernando et al. (2020)	-	2	1
JESC Pryzant et al. (2017)	-	2	1
KFTT Neubig (2011)	v1.0	2	1
ASPEC Nakazawa et al. (2016)	-	2	1

7. By convention, we use `sq` for Albanian, and don't distinguish dialects like Gheg (`aln`) and Tosk (`als`)

8. We use `ber` as the code for Tamazight, after consultation with Tamazight speakers opining that the dialect distinctions are not significant. Other resources use the individual codes like `tzm` and `kab`.

9. We use the macrocode `qu` for Quechua. In practice, this seems usually to be a mix of the Ayacucho and Cusco dialects. Other resources, like NLLB, may use the dialect code, e.g. `quy` for Ayacucho Chanka. The same is true for a few other macro codes, like `ff` (Macro code for Fulfulde, whereas other sources may use e.g. `fuv`.)

10. Really, there are notes that can be made about almost any code, from the well-accepted conventions like `zh` for Mandarin, to many dialectical notes, like which variant of Hmong really is the `hmn` data? But The above ones are made specifically for ones where we are aware of other datasources floating out there that use different conventions.

A.7 Multiway Data Details

On creating the multiway data described in Section 2.1.3.4, we obtain a dataset with 11.9B sentence pairs across 19.7k language pairs. In Table A.4, we list the combined number of sentence pairs for each target language.

A.8 Model Training Details

MT Model Training We train models of various sizes: a 3B, 32-layer parameter model,[1] a 7.2B 48-layer parameter model and a 10.7B 32-layer parameter model. We describe the specifics of the model architecture in Table A.5.

We share all parameters of the model across language pairs, and use a Sentence Piece Model (SPM) (Kudo, 2018) with 256k tokens shared on both the encoder and decoder side. We train

[1]Here and elsewhere, 'X-layer' means X encoder layers and also X decoder layers, for a total of 2X layers.

the SPM model on upto 1M sentence samples of the sentences in the sentence-level version of MADLAD-400, supplemented by data from the languages in the parallel data used to train MT models when not available in MADLAD-400 with a temperature of $T = 100$ and a character coverage of 99.9995%.

Each input sentence has a <2xx> token prepended to the source sentence to indicate the target language (Johnson et al., 2017). We use both supervised parallel data with a machine translation objective and the monolingual MADLAD-400 dataset with a MASS-style (Song et al., 2019) objective to train this model. Each of these objectives is sampled with a 50% probability. Within each task, we use the recently introduced UniMax (Chung et al., 2023) sampling strategy to sample languages from our imbalanced dataset with a threshold of $N = 10$ epochs for any particular language.

We used a square root learning rate decay schedule over the total number of training steps, starting at 0.01 and ending at X, as well as the AdaFactor optimizer with factorized=False and 10k warmup steps. We note that for the 10.7B model we use a dropout probability of $p = 0.2$ instead of $p = 0.1$ in order to mitigate overfitting to low-resource languages.

Language Model Training We follow the same training schedule and model configurations from Garcia et al. (2023). In particular, we consider 8B decoder-only models, following the same model hyperparameters as previous work (Chowdhery et al., 2022; Garcia et al., 2023). We train these models using a variant of the UL2 objective (Tay et al., 2022b) adapted for decoder-only models, and use the same configuration as previous work (Garcia et al., 2023; Orlanski et al., 2023). We point the reader to these papers for a detailed overview of the training process, and include basic architectural details in Table A.5. We use the same SPM trained for the MT models.

We use a square root learning rate decay schedule over 500k steps, starting at 0.01 and ending at 0.001, as well as the AdaFactor optimizer with factorized=False with 10k warmup steps. We describe the evaluation setup in Section 2.1.4.3.

Table A.5: Architecture and training details for the various models we train in this work using MADLAD-400.

Model	SPM Size	Training Objective	# Attn Heads	# Enc. Layers	# Dec. Layers	Model Dim	Hidden Dim	Batch Size	Max. Seq. Length	Training Steps
3B MT Model	256k	MT+MASS	16	32	32	1024	8192	4096	256	1M
7.2B MT Model		MT+MASS	16	32	32	2048	8192	4096	256	500k
10.7B MT Model		MT+MASS	32	48	48	2048	16384	4096	256	250k
8B LM		UL2	16	N/A	32	4096	16384	1024	1024	500k

A.9 Languages Evaluated

In Table A.6, we list the languages for which we evaluate the models trained as described in Sections 2.1.4.1 and 2.1.4.2.

Table A.6: Languages on which we evaluate the trained models for each multilingual evaluation set.

Datasets	**Languages Evaluated**
WMT	cs, de, es, fi, fr, gu, hi, kk, lv, lt, ro, rs, es, tr, zh

Continued on next page

Table A.6 continued from previous page

Datasets	Languages Evaluated
Flores-200	ac_Arab, ace, af, am, ar, arz, as, awa, ay, az, ba, ban, be, ber, bg, bho, bjn_Arab, bjn, bm, bn, bo, bs, bug, ca, ceb, ckb, crh_Latn, cs, cy, da, de, din, dyu, dz, ee, el, eo, es, et, eu, fa, ff, fi, fil, fj, fo, fon, fr, fur, ga, gd, gl, gn, gu, ha, hi, hne, hr, ht, hu, hy, id, ig, ilo, is, it, he, ja, jv, ka, kac, kbp, kg, kk, km, kmb, kn, ko, kr_Arab, kr, ks_Deva, ks, ky, lb, lg, li, lij, lmo, ln, lo, lt, ltg, lus, lv, mag, mai, mg, mi, min, mk, ml, mn, mni, mr, ms, mt, my, ne, nl, nn, no, nso, nus, ny, oc, om, or, pa, pag, pap, pl, ps, pt, quy, rn, ro, ru, rw, sa, sc, scn, sd, sg, shn, si, sk, sl, sm, sn, so, sq, sr, ss, st, su, sv, sw, szl, ta, taq_Tfng, taq, te, tg, th, ti, tk, tn, tr, ts, tt, ug, uk, ur, vec, vi, war, wo, xh, yi, yo, zh_Hant, zh, zu

Continued on next page

Table A.6 continued from previous page

Datasets	Languages Evaluated
NTREX	af, am, ar, az, ba, be, bem, bg, bn, bo, bs, ca, ckb, cs, cy, da, de, dv, dz, ee, el, en_GB, en_IN, es, es_MX, et, eu, fa, fa_AF, ff, fi, fil, fj, fo, fr, fr_CA, ga, gl, gu, ha, hi, hmn, hr, hu, hy, id, ig, is, it, iw, ja, ka, kk, km, kn, ko, ku, ky, lb, lo, lt, lv, mey, mg, mi, mk, ml, mn, mr, ms, mt, my, nd, ne, nl, nn, no, nso, ny, om, pa, pl, ps, pt, pt_PT, ro, ru, rw, sd, shi, si, sk, sl, sm, sn, so, sq, sr, sr_Latn, ss, sv, sw, ta, te, tg, th, ti, tk, tn, to, tr, tt, ty, ug, uk, ur, uz, ve, vi, wo, xh, yo, yue, zh, zh_Hant, zu
Gatones	ady, ak, as, av, ay, ba, ban, bbc, bci, ber_Latn, bew, bho, bm, bo, ce, chr, ckb, cv, doi, dv, dyu, dz, ee, ff, gn, gom, ilo, iso, kl, kri, lg, ln, lus, mad, mai, meo, min, mni, nso, om, or, qu, quc, rw, sa, sg, skr, ti, tiv, tk, ts, tt, ug, wo, yua, zza

A.10 Results Details

In Tables A.7, A.11, A.8, A.9 and A.10 we list the WMT, NTREX, Gatones, Flores-200 and Flores-200 (direct pairs) chrf and SacreBLEU scores respectively by language pair along with the model checkpoints.

Table A.8: Evaluation scores on the GATONES test set used by Bapna et al. (2022) (depicted as <bleu> / <chrf>) for the MT models and language models described in Section 2.1.4.1 and Section 2.1.4.2.

	NTL (Bapna et al. (2022))		**MT-3B**	**MT-7.2B**	**MT-10.7B**	**LM-8B**			
	1.6B	6.4B				0-shot	1-shot	5-shot	10-shot
ady_en	- / 53.7	- / 54.8	32.7 / 58.8	34.2 / 59.8	33.9 / 60.1	1.5 / 12.4	18.9 / 50.2	25 / 54.2	21.1 / 52.1
ak_en	- / 31.7	- / 36.3	6.9 / 27	10.4 / 29.9	9.3 / 29	0.1 / 5.5	3.6 / 15.5	1.8 / 18.6	1.9 / 16.6
as_en	- / 52.9	- / 58.6	30.6 / 55.6	32.1 / 56.8	33.8 / 58	1 / 9.7	17.4 / 44.5	19.4 / 47.4	21.1 / 48.6
av_en	- / 48.1	- / 48.1	19.7 / 47.8	21.3 / 50.6	21.5 / 50.9	0.3 / 6.1	11.6 / 32.5	10.3 / 40.2	11.1 / 40.6
ay_en	- / 32.3	- / 30.8	9.1 / 28.7	11.3 / 29	11.3 / 30.1	0.2 / 5.6	1.4 / 13.7	5.9 / 23.1	5.5 / 22.6
ba_en	- / 42.3	- / 45.8	4.5 / 23.3	5.4 / 24	6.9 / 27.2	0.2 / 2.9	8.7 / 30.3	7.6 / 35	7.7 / 34.6
ban_en	- / 32.2	- / 35.4	10.1 / 30.9	10.3 / 31.2	10.9 / 32.4	0.2 / 6.6	2.2 / 21.2	4.7 / 24.6	4.7 / 22.7
bbc_en	- / 40.9	- / 44	12.4 / 36.6	12.1 / 36.7	13.3 / 38	0.2 / 6.6	5.6 / 27.5	7.7 / 29.4	7.9 / 28.4
bci_en	- / 11.6	- / 10.8	2.7 / 19.7	3.6 / 20.4	3.9 / 21.6	0 / 4	0.6 / 16.8	0.5 / 9.5	1.9 / 15.3
bew_en	- / 49.8	- / 51.8	25.3 / 50.1	26.6 / 51.4	26.7 / 51.7	0.7 / 10.8	20.3 / 44.8	23.8 / 49.2	21.3 / 48.3
bho_en	- / -10000	- / -10000	28.6 / 55.8	25.6 / 52.9	29.3 / 56.3	0.6 / 7.4	17.7 / 45.6	20 / 49.1	21.6 / 49.7
bm_en	- / 27.6	- / 36.3	11.9 / 32.8	10.6 / 31.7	12.6 / 33.9	0.1 / 4.2	0.1 / 11.5	1.2 / 11	1.2 / 14.4
ce_en	- / 44.6	- / 44.9	11.4 / 36	12.5 / 37	13.5 / 37.3	0.6 / 7.5	11.4 / 36.4	9.4 / 33.2	15.1 / 41.3
chr_en	- / 16.3	- / 15.6	1.1 / 5.9	0.6 / 8	0.9 / 7.2	0 / 0.6	0.2 / 12.8	0.3 / 12.6	0.1 / 10
cv_en	- / 39.9	- / 46.3	19 / 42.8	21.2 / 44.9	22.9 / 46.6	0.3 / 5.7	7.5 / 24.6	13 / 38	12.2 / 36.3
doi_en	- / 57.4	- / 63.1	15.3 / 43.1	16 / 42.9	18 / 45	0.2 / 4.7	6.9 / 31.9	8.8 / 35.2	6.1 / 36.4
dv_en	- / 37	- / 45.2	20.1 / 45.3	21.4 / 47.1	22.5 / 47.7	0.8 / 8.9	11.4 / 35.5	13.2 / 38.1	14.6 / 41.5
dyu_en	- / 23.9	- / 23.5	7.6 / 28	6.7 / 27.7	8.6 / 30	0 / 4.6	2 / 14.1	2.3 / 16.2	1.3 / 15.1
ee_en	- / 29.1	- / 33.5	8.6 / 25.8	11.1 / 28.9	10.4 / 28.3	0.1 / 6	2 / 17.6	2.1 / 18.9	3.1 / 16.4
en_ady	- / 21.8	- / 28.2	0.6 / 8.8	2 / 14.2	1.4 / 11.6	0.1 / 0.6	2.1 / 18.9	0.4 / 11.2	1.7 / 28.6
en_ak	- / 32.1	- / 34.3	2.2 / 17.9	4.1 / 21.9	3.9 / 21.2	0.3 / 5.9	0.1 / 3.5	0.1 / 4.5	0 / 3.1
en_as	- / 36.3	- / 36.7	8.4 / 36.7	8.4 / 36.9	8.7 / 38	0.1 / 1.8	0.7 / 17.6	0.7 / 18.4	0.9 / 19.1
en_av	- / 26.1	- / 28.1	1.7 / 17.5	1.8 / 21.5	1.9 / 20.4	0.1 / 0.5	0.1 / 4.5	0.5 / 15.8	0.1 / 5.9
en_ay	- / 30	- / 30.5	1.4 / 17	2.4 / 21.7	2.1 / 21.6	0.2 / 6.2	0.5 / 9.3	0.2 / 4.7	0 / 3.4
en_ba	- / 37.6	- / 38.4	2.3 / 22.1	3.6 / 25.3	4 / 26.8	0.1 / 0.9	0.2 / 8.8	0.2 / 9.3	0.1 / 5.7
en_ban	- / 31.8	- / 33.1	3.9 / 30.8	3.7 / 30.9	4 / 31.5	0.2 / 8.5	0.7 / 14.1	0.1 / 4.4	0.1 / 8.5
en_bbc	- / 34.9	- / 35.4	1.9 / 26.8	2.6 / 28.8	3.1 / 29.5	0.1 / 4.4	0.4 / 12.5	0.1 / 8.1	0.1 / 7.1
en_bci	- / 14.1	- / 14.7	1.5 / 12.7	2 / 13.8	2.4 / 14.9	0.1 / 4.8	0 / 2.2	1.1 / 8.8	0 / 1.7
en_bew	- / 45.7	- / 46	7.9 / 35.6	12.6 / 44.9	12.7 / 45.5	0.1 / 6.2	7.5 / 35.2	10.4 / 41.9	11.7 / 43.9
en_bho	- / 0	- / 40.6	12.8 / 39.5	12 / 38.4	12.7 / 40.1	0.2 / 3.8	2.9 / 24.6	4.1 / 28.8	5.8 / 30.5
en_bm	- / 28.7	- / 34.3	6.5 / 28.5	6.1 / 27.2	7.1 / 30.5	0.2 / 5.5	0.1 / 4.8	0 / 2.6	0 / 2.1
en_ce	- / 23.5	- / 23.7	3.4 / 9	7.5 / 22.9	1.9 / 15.2	0.2 / 1.3	0.8 / 8.7	0 / 1.2	0.1 / 3.5
en_ckb	- / 0	- / 41.6	7.5 / 38.7	7.9 / 38.4	6.7 / 35.4	0.1 / 0.6	0.2 / 8.2	0.2 / 9	0.2 / 11.1
en_cv	- / 28.1	- / 32.1	3.1 / 25.6	3.1 / 25.6	3.1 / 25.8	0.1 / 0.9	0.1 / 5	0.1 / 5.5	0 / 4.5
en_doi	- / 29.9	- / 25.5	0.7 / 0.8	0.7 / 0.8	0.7 / 0.8	0.1 / 0.2	0.3 / 10.1	0.3 / 11	0.1 / 5.9
en_dv	- / 43.2	- / 43.7	2.9 / 38	3.5 / 40.2	3.8 / 40.8	0.1 / 0.2	0.6 / 0.4	0.6 / 0.5	0.5 / 0.4
en_dyu	- / 12.7	- / 20.4	1.9 / 19.4	2.3 / 19.6	2.2 / 20.1	0.1 / 4	0.4 / 11.2	0 / 4.6	0 / 1.5
en_ee	- / 35.9	- / 39.2	4.1 / 23.7	6.1 / 27.1	6 / 26.6	0 / 4.7	0.1 / 4.2	0 / 2.9	0 / 2.1
en_ff	- / 31.1	- / 32.3	3 / 21.4	3.3 / 21.3	3.7 / 22.9	0.1 / 5.5	0 / 12	0 / 1.8	0 / 3.3

en_gn	- / 24.3	- / 32.2	6.3 / 30.8	6.4 / 30.3	6.4 / 31.4	0.2 / 6.7	0.2 / 14.6	0.1 / 4.4	0 / 2.5
en_gom	- / 1.1	- / 39.1	1.2 / 20.1	2.1 / 23.5	4.3 / 28.2	0.1 / 0.7	0.6 / 15.3	0.5 / 16.8	1.2 / 20
en_ilo	- / 49.7	- / 52.4	13.9 / 37.8	16.2 / 41.3	16.9 / 42.8	0.3 / 7.1	2.5 / 23.6	3.4 / 25.7	2.1 / 24.8
en_iso	- / 16.5	- / 30.5	2.1 / 16.3	2.4 / 18.2	3.1 / 17.4	0.2 / 4.3	0 / 3.2	0.1 / 7.2	0 / 4.2
en_kl	- / 0	- / 23.1	2.6 / 29.5	3.6 / 33	3.4 / 33.6	0.2 / 4.6	0.4 / 9	0.1 / 4.7	0 / 4.8
en_kri	- / 32.7	- / 34.9	2.6 / 22.2	2.3 / 18.2	3.4 / 22.7	0.3 / 5.9	0.7 / 11.3	0.2 / 8.7	0.7 / 16.4
en_lg	- / 35.9	- / 38	1.7 / 23.6	1.6 / 25.8	2 / 27.4	0.1 / 5.3	0 / 3.6	0 / 3.6	0 / 2.6
en_ln	- / 12.7	- / 33.9	0.3 / 11.7	0.4 / 13.6	0.6 / 15.2	0 / 5.8	0 / 2.7	0 / 2.9	0 / 1.7
en_lus	- / 16.9	- / 39.3	7.7 / 31.1	10.6 / 35.6	10.2 / 34.7	0.2 / 5.6	0.1 / 4.8	0.1 / 4.7	0.1 / 4.2
en_mad	- / 29.8	- / 30.2	2.1 / 23.8	3.3 / 25.9	4 / 25.5	0.1 / 6.8	0.1 / 4.6	0.1 / 6.4	0.1 / 3.2
en_mai	- / 34.1	- / 37.6	12.7 / 42.4	13.2 / 41.2	14 / 43.2	0.1 / 0.8	1.6 / 18.2	3.4 / 28.9	3.4 / 29.5
en_meo	- / 50.9	- / 50.7	39.3 / 63	46.3 / 68.9	44.1 / 67.6	2.2 / 13.2	47.4 / 70.4	46.5 / 70.3	48.3 / 70.8
en_min	- / 51.8	- / 56.1	10.7 / 44.7	12.9 / 47.2	10 / 45	0.1 / 7.3	5.7 / 35.9	4.7 / 34.7	6.3 / 39.9
en_mni	- / 35.7	- / 38.2	7.1 / 38.9	7.9 / 39.4	7.9 / 40.6	0.1 / 0.3	0 / 1.7	0 / 1.7	0 / 0.6
en_nso	- / 45	- / 41.6	11.1 / 34.4	14.4 / 38.2	13.4 / 37.3	0.2 / 4.5	0.1 / 5.9	0.1 / 5.6	0.1 / 5
en_om	- / 36	- / 39.1	1 / 20	1.7 / 26.9	1.9 / 26.4	0.1 / 5.8	0 / 4.4	0 / 2.6	0 / 2.4
en_qu	- / 29.8	- / 33.1	1.5 / 19.5	2.3 / 21.1	1 / 22.4	0 / 6	0.1 / 6.7	0 / 2.2	0 / 1.1
en_quc	- / 22.8	- / 22.9	0.5 / 10.1	1 / 12.7	1.4 / 14.2	0.2 / 5.9	0 / 5.7	0 / 3.2	0 / 4.5
en_sa	- / 26.9	- / 28.4	1.3 / 24	1.5 / 26.9	1.8 / 27.8	0 / 1.2	0 / 5.3	0 / 5.7	0 / 6.2
en_sg	- / 12.7	- / 20.7	0.4 / 18.5	0.5 / 18.2	0.5 / 18.5	0 / 4.7	0 / 1.3	0.1 / 22.2	0 / 1.3
en_skr	- / 32.8	- / 31.3	1.3 / 21.9	1.5 / 24.9	1.9 / 25	0 / 0.1	0.1 / 7.9	0 / 6	0 / 4.8
en_ti	- / 19.9	- / 21.1	1.3 / 12.8	1.5 / 15.6	1.7 / 15.8	0.1 / 0.3	0 / 0.6	0 / 0.6	0 / 0.4
en_tiv	- / 13.7	- / 23	1 / 16.2	1.3 / 18.5	1 / 17.9	0.2 / 4.7	0.4 / 9.3	0 / 2.3	0 / 3.7
en_ts	- / 17.4	- / 45.5	6.4 / 33.6	8.9 / 37.4	7.4 / 35.7	0.1 / 5.2	0.1 / 4.1	0.1 / 5.8	0.1 / 5
en_yua	- / 30.6	- / 31.6	3.5 / 19	1.4 / 17.2	3 / 19.5	0.2 / 6.2	0.6 / 9.9	0.1 / 4	0.1 / 7
en_zza	- / 22.1	- / 22.6	1.7 / 19.8	1.1 / 18	1.3 / 17.6	0 / 3.6	0 / 8.9	0 / 2.6	0 / 1.6
ff_en	- / 33.7	- / 41.2	7.8 / 26.8	8.3 / 26.5	9.4 / 29.2	0.1 / 4.4	0.4 / 12.5	1.1 / 15.4	1.3 / 13.6
gn_en	- / 26.7	- / 38.9	12.9 / 35.5	12.7 / 35.2	13.7 / 37	0.2 / 6	1.5 / 16.7	3.9 / 22.2	2.6 / 19.8
gom_en	- / 50.7	- / 55.5	20.5 / 45.7	20.4 / 45	21 / 45.7	0.1 / 1.2	10.5 / 33.2	16.6 / 42.1	15.7 / 41
ilo_en	- / 50.8	- / 43.4	17.3 / 44.2	28.6 / 52.7	29.6 / 52.8	0.2 / 5.7	13.7 / 38.9	20.1 / 42.7	18 / 42.1
iso_en	- / 17.9	- / 29.4	5.8 / 24.3	7.7 / 25.3	8.2 / 26.1	0.1 / 4.5	1.1 / 16	1.6 / 14.4	2.3 / 17.2
kri_en	- / 48.7	- / 56.5	13.2 / 37	15.6 / 39.5	17.3 / 41.5	0.1 / 6	10.5 / 31.3	7.5 / 31.1	10.5 / 33.2
lg_en	- / 32.6	- / 37.9	6.3 / 24.1	8.9 / 26.7	8.2 / 27.1	0 / 3.9	1.9 / 18.9	2.5 / 18.3	3.3 / 17.3
ln_en	- / 28.2	- / 30.4	0.8 / 13.1	0.9 / 14.2	0.8 / 14.8	0.1 / 5.1	0.3 / 15.2	1 / 15	0.9 / 12.7
lus_en	- / 22.6	- / 34.5	13 / 35.7	18.4 / 40.9	17.9 / 41.2	0.6 / 12	7.5 / 31	12.1 / 33.6	10.8 / 33.4
mad_en	- / 44.9	- / 50.5	8.5 / 34	10.4 / 34.3	11.9 / 36.7	0.1 / 5.9	3.8 / 22.5	3.9 / 24.9	5 / 26.8
mai_en	- / 57.8	- / 61.6	30.9 / 57	29.2 / 55.7	33.3 / 58.8	0.9 / 9.3	15.3 / 45.1	20.5 / 50.1	20.3 / 49.2
meo_en	- / 66.1	- / 67.9	44.4 / 66.4	46.2 / 67.8	45.5 / 67.5	0.9 / 9.4	38.7 / 61.5	39.6 / 63.3	41.1 / 63.7
min_en	- / 58.8	- / 62.4	26.2 / 52.4	25.3 / 50.8	29 / 55.1	0.4 / 7.7	16.7 / 39.3	20.9 / 49.1	21.7 / 49.5
mni_en	- / 54.8	- / 56.4	32.2 / 55.9	31 / 54.9	33.4 / 57	0 / 1.9	1.2 / 17.5	2.6 / 20.4	2.7 / 20.5
nso_en	- / 45.8	- / 51.3	17.6 / 40.2	21.1 / 43.2	21.1 / 43.4	0.2 / 6.1	5.7 / 23.9	7.6 / 28.7	6.1 / 27.3
om_en	- / 32.1	- / 38.1	5.7 / 25.8	9.4 / 30.6	9.4 / 31.3	0.2 / 6.2	0.8 / 18.9	2.8 / 19	2.3 / 20.2
qu_en	- / 29.9	- / 32.5	6.7 / 27.2	7.3 / 27.6	7.5 / 28.2	0.1 / 6.1	2.5 / 18.1	1.8 / 13.7	2.3 / 18.8
quc_en	- / 24.4	- / 26.7	1.9 / 17.6	2 / 15.9	2.8 / 19.3	0.1 / 7.1	1.3 / 11.4	1.6 / 13.9	0.9 / 12.3
sa_en	- / 43.5	- / 46.3	15.1 / 39	17.3 / 40.4	14.5 / 37.9	2.6 / 18.1	7 / 35.8	7.4 / 34.8	9.8 / 37.3
sg_en	- / 13.4	- / 13.8	0.5 / 12.5	0.6 / 13.5	0.7 / 14.1	0 / 3.2	0 / 17	0.2 / 9.1	0.2 / 12.1
skr_en	- / 44.4	- / 48.4	9 / 35.5	12.5 / 38.8	12.5 / 39.1	0.3 / 8.5	6.5 / 28.5	9 / 33.2	8.5 / 34.3
ti_en	- / 37.9	- / 44.2	11.7 / 34	15.2 / 38.2	15.2 / 38.2	0.3 / 7.1	5.3 / 22.3	7.1 / 27.4	7 / 26.3
tiv_en	- / 14.1	- / 15	2.7 / 14	2.9 / 14.8	3 / 15.4	0.1 / 3.8	0.9 / 12.3	1.1 / 11.8	0.9 / 12.2
ts_en	- / 25.5	- / 43	14.3 / 35.7	18.5 / 40.5	17.8 / 40	0.3 / 7.5	1 / 16.3	7.6 / 25.9	8.7 / 27.8

yua_en	- / 34.6	- / 40.7	6 / 27	7.1 / 28.3	9.7 / 30.9	0.2 / 6.9	0.8 / 11.3	3.3 / 20.2	4.4 / 19.2
zza_en	- / 23.4	- / 23.2	4.6 / 22.4	5.1 / 22.8	5.2 / 24.2	0.2 / 10.6	1.9 / 18.9	3.3 / 21.3	2.3 / 20.6
xx2en	- / 37.2	- / **41.2**	13.3 / 34.6	**15.5** / 36.5	14.8 / 36.0	0.3 / 6.5	6.6 / 25.4	8.3 / 28.1	8.4 / 28.4
en2xx	- / 28.5	- / **33.1**	4.5 / 23.9	**5.5** / 26.4	5.4 / 26.2	0.2 / 4.2	1.7 / 10.5	1.7 / 9.9	1.8 / 9.4
Average	- / 32.9	- / **37.2**	8.9 / 29.3	**10.5** / 31.5	10.1 / 31.1	0.3 / 5.4	4.2 / 18.0	5.0 / 19.0	5.1 / 18.9

Table A.9: Evaluation scores on en-centric Flores-200 pairs (depicted as <bleu> / <chrf>) for the MT models and language models described in Section 2.1.4.1 and Section 2.1.4.2 compared against NLLB-54B. All metrics are computed with the sacrebleu reference implementation.

	NLLB	**MT-3B**	**MT-7.2B**	**MT-10.7B**	**LM-8B**			
					0-shot	1-shot	5-shot	10-shot
en_ace_Arab	0.4 / 21.1	1.4 / 26.6	1.4 / 25.3	1.3 / 26.9	0 / 0.1	0 / 0.5	0 / 2.9	0 / 2.2
en_ace	9.3 / 41.4	8.3 / 39.6	6.5 / 36.6	7.6 / 39.1	0.1 / 3.7	0.9 / 15.3	0.8 / 21.1	0.7 / 22.4
en_acm	10.3 / 35.4	0.6 / 0.8	0.4 / 0.8	0.6 / 0.8	0 / 0.4	4.8 / 34.6	8.3 / 43.3	6.1 / 41.1
en_acq	13.6 / 46.8	0.6 / 0.8	0.4 / 0.7	0.6 / 0.8	0.1 / 1.2	5.9 / 36.1	10.1 / 43.5	9.7 / 44
en_aeb	12 / 42.1	0.6 / 0.8	0.6 / 0.8	0.6 / 0.8	0.1 / 1.2	1.4 / 20.3	4.1 / 35.4	4.9 / 37.2
en_af	38.4 / 66.6	38.8 / 66.1	40.9 / 67.6	40.4 / 67.3	0.7 / 10.5	31.2 / 62.2	36.9 / 64.1	38.6 / 64.3
en_ajp	20.5 / 56	0.7 / 1.3	0.6 / 1.3	0.7 / 1.3	0 / 1	9.5 / 41.9	9.3 / 43.1	9.7 / 44.5
en_am	15 / 42.2	8.5 / 31.2	8.8 / 32.2	9.1 / 32.5	0 / 0.2	0.1 / 0.6	0.2 / 0.7	0.2 / 0.7
en_apc	20.4 / 54.9	0.8 / 1.2	0.6 / 1.2	0.7 / 1.1	0 / 1	6 / 37.7	5.5 / 39	4.4 / 40
#N/A en_ar_MA	11.7 / 43.5	0.4 / 0.9	0.3 / 0.8	0.3 / 0.8	0 / 1.1	0.2 / 5.5	1 / 24.7	0.2 / 15.3
en_ar	29.7 / 60.6	27.3 / 57.9	28.2 / 58.4	27.7 / 58.2	0.1 / 1.2	12.6 / 44	17.5 / 50.2	18 / 50.6
en_ars	23.2 / 54.2	0.4 / 0.6	0.3 / 0.6	0.4 / 0.6	0.1 / 1.1	11.6 / 42.3	15.1 / 48.3	16.2 / 49.4
en_arz	16.7 / 50.9	16.6 / 48.7	15.1 / 47	16.4 / 48.9	0.1 / 1.2	2.6 / 27	3.4 / 34.8	6 / 38.6
en_as	7.9 / 40.3	7.3 / 37.3	8.1 / 38.9	8 / 39.2	0 / 0.9	0.2 / 12.9	0.7 / 21.8	0.6 / 21.3
en_ast	29.8 / 59.6	1.1 / 2.1	2 / 13.6	1.1 / 2	0.2 / 7.8	6.4 / 38.4	9.7 / 44.6	11.2 / 45.9
en_awa	18.5 / 50.7	10.3 / 43.8	9.5 / 42.3	10.2 / 44.4	0 / 0.3	0.8 / 16.5	7.2 / 38	5 / 38.6
en_ay	4 / 34.5	0.8 / 16.7	1.1 / 19.5	1.4 / 20.9	0 / 2.6	0 / 2	0 / 7	0 / 4.1
en_az	14.1 / 47.1	12.5 / 44.2	13.5 / 45.3	13.9 / 46.4	0.1 / 5.4	8.4 / 38.8	9.7 / 42.8	10.6 / 43.7
en_azb	1.3 / 27.8	0.2 / 0.6	0.2 / 0.6	0.2 / 0.6	0 / 0.1	0 / 0.8	0 / 6.9	0 / 5.6
en_ba	18.7 / 51	5.9 / 28.7	7.8 / 31.7	10.4 / 36.8	0 / 0.3	0.3 / 10.8	0.3 / 19.8	0.2 / 16.4
en_ban	15.3 / 48.4	13.1 / 45.5	11.1 / 42.4	13.4 / 45.9	0.1 / 5.5	0.4 / 10.3	0.7 / 21.7	1.7 / 30.4
en_be	14.4 / 45.7	12.3 / 40.5	13.5 / 42.1	13.6 / 42.8	0.1 / 2.6	11.5 / 40.9	11.5 / 41.5	11.1 / 41.6
en_bem	10.5 / 42.3	1.6 / 2.4	2.7 / 17.7	1.6 / 2.7	0.2 / 4.4	0 / 4.9	0.1 / 10.1	0.1 / 7.9
en_ber	8.1 / 34.3	6.6 / 31	6.2 / 30.6	7.5 / 33.4	0 / 0.2	0 / 0.5	0 / 2	0 / 1
en_bg	41.6 / 67.1	43.5 / 68.9	43.9 / 69.2	44.1 / 69.3	0.7 / 5.7	32.2 / 60.2	33 / 61.4	33.4 / 62.2
en_bho	17.3 / 45.2	13.4 / 41	11.7 / 38.8	13.8 / 42.1	0.1 / 2.1	1.4 / 19.5	4.6 / 31.7	4 / 31.2
en_bjn_Arab	1.1 / 20.3	3.3 / 30.5	2.9 / 28.6	3.6 / 31	0 / 0	0 / 1.1	0 / 4.4	0 / 1.5
en_bjn	19.2 / 51.8	15.8 / 49.6	13.3 / 45.6	15.8 / 49.8	0 / 2.4	1 / 17.2	4.9 / 36.3	3 / 34.9
en_bm	6.8 / 32.9	4.9 / 27.8	4.3 / 26.8	5.7 / 30.1	0 / 3.1	0 / 6.3	0 / 11.2	0.1 / 7.8
en_bn	19.4 / 54.2	14 / 47.2	14.8 / 48.2	14.8 / 48.3	0.1 / 3.7	4.5 / 35	8.2 / 44.4	7.1 / 44.3
en_bo	0.8 / 38.6	13.8 / 26.4	15.9 / 29.4	15.9 / 30.2	0 / 0.1	0 / 0	0 / 1.4	0 / 1
en_bs	33.2 / 61.4	33.4 / 61.9	34 / 62.3	34.6 / 62.9	0.3 / 5.9	24 / 53.6	27.6 / 56.9	26.9 / 57.5
en_bug	6.5 / 38.3	5.8 / 36.6	5.3 / 34.7	5.8 / 37.2	0.1 / 4.3	0 / 2.2	0 / 6.7	0.1 / 8.4
en_ca	43.9 / 66.8	45.8 / 67.5	46.4 / 68	46.7 / 68.4	0.9 / 11.3	35.9 / 60.7	36.3 / 61.8	37.1 / 61.9
en_ceb	30.3 / 59.6	7.4 / 26.5	8.3 / 27.6	9.4 / 29.4	0.2 / 3.3	9.1 / 39.8	9.3 / 39.4	10.3 / 40.7
en_cjk	2.7 / 28.3	1.3 / 2.4	1.2 / 16.6	1.4 / 3.8	0 / 3.3	0 / 1.3	0.1 / 6.2	0 / 6.2
en_ckb	13.2 / 52.5	9.2 / 44.9	9.8 / 45	9.4 / 44.4	0 / 0.1	0.1 / 5.6	0.2 / 14	0.1 / 13.5
en_crh_Latn	16.8 / 51	13.2 / 47.3	10.9 / 44.1	13.8 / 48.4	0.1 / 3.6	0.4 / 15.2	1.1 / 27.5	1.3 / 28.2
en_cs	33.6 / 59.8	34.8 / 61.3	34.7 / 61.1	35.6 / 61.5	0.2 / 5.6	22.5 / 52.1	26.2 / 54.5	25.9 / 54.6
en_cy	51.6 / 72.5	41.7 / 62.4	45.7 / 66.4	44.7 / 65.6	0.1 / 5.1	1.5 / 15.2	8.8 / 38.4	4.9 / 35.6

en_da	44.5 / 68.4	48 / 71.1	47.9 / 71.1	48.3 / 71.3	0.6 / 8.5	40.3 / 65.3	41 / 65.4	42.1 / 66.8
en_de	39.6 / 65.2	41.5 / 67.1	42 / 67.5	42.2 / 67.4	0.5 / 10.1	27.6 / 56.3	28.4 / 57.4	29.1 / 58
en_din	3.8 / 26.8	3.4 / 25.1	3.2 / 24.7	3.3 / 25.9	0.1 / 3.4	0 / 2.9	0 / 10.6	0 / 0.9
en_dyu	1.7 / 19.8	2 / 21	2.3 / 21.1	2.1 / 22.1	0 / 2.7	0 / 2.7	0 / 7.4	0 / 5.8
en_dz	0.6 / 45.4	34.2 / 42.5	32.3 / 41	37 / 45.1	0 / 0	0 / 1	0 / 3.4	0 / 2.4
en_ee	13.2 / 41.9	5.2 / 26.9	6.8 / 29.2	5.9 / 28.4	0 / 3.2	0.1 / 3.4	0.1 / 7.5	0.1 / 5.9
en_el	28 / 54.2	29.2 / 55.9	29 / 55.8	29.5 / 56.3	0.3 / 2.4	19 / 46.5	21.5 / 48.8	20.9 / 48.9
en_eo	35.3 / 63.9	33.6 / 62.6	34.1 / 63	34.3 / 63.7	0.2 / 5.4	5.1 / 32.4	12.6 / 47.2	9.4 / 44.8
en_es	28.1 / 56	28 / 56.4	28.4 / 56.6	28.5 / 56.7	0.3 / 8.9	22.6 / 50.6	23.6 / 52.1	23.7 / 52.2
en_et	27.1 / 59.9	29 / 61.3	28.5 / 61.2	29.8 / 61.8	0.2 / 6	20.4 / 54.1	22.7 / 55.5	23.1 / 55.5
en_eu	17.2 / 54.9	18.3 / 55.5	18.6 / 55.7	19.3 / 57.2	0 / 5.5	8.3 / 45.5	12.5 / 51.1	12.5 / 51.3
en_fa_AF	27.8 / 55.4	0.5 / 1	0.4 / 0.9	0.5 / 0.9	0 / 0.1	14.7 / 44.4	21.6 / 48.4	19.4 / 48.4
en_fa	24.9 / 53.5	24.8 / 51.2	25.3 / 51.4	25.1 / 51.7	0.1 / 0.8	15.3 / 44.7	20.7 / 49.5	20.1 / 49.2
en_ff	3.9 / 0	2.2 / 23.4	2.5 / 24	2.5 / 25.1	0 / 3.6	0 / 20	0 / 3.8	0 / 7.8
en_fi	26.2 / 59.4	26.9 / 61.7	27.3 / 62	27.7 / 62.6	0.1 / 7.2	19.1 / 52.5	20 / 53.9	19.9 / 54.2
en_fil	36.5 / 62.3	28.6 / 56.5	28.8 / 56.3	28.2 / 56.5	0.3 / 6.4	29 / 55.6	31.1 / 58.3	31.1 / 57.9
en_fj	20.1 / 48.8	9 / 32.5	11.1 / 35.6	10.4 / 34	0.1 / 4	0.1 / 13.9	0.1 / 9.4	0.1 / 8.3
en_fo	26.3 / 51.8	16.9 / 40	18.5 / 41.5	17.7 / 41.2	0.1 / 7.3	5.6 / 29.4	6.1 / 31.1	7.7 / 33.2
en_fon	3.1 / 23.5	1.6 / 15.6	1.9 / 16	1.6 / 15.9	0 / 0.3	0 / 1	0 / 1.9	0 / 1.6
en_fr	51.4 / 71.4	52.5 / 72.1	53.4 / 72.5	53.2 / 72.5	0.4 / 9.2	36.1 / 61.4	39.1 / 63.5	38.8 / 63.2
en_fur	34 / 58.7	28.8 / 55.8	24.7 / 51.4	28.8 / 56.1	0.1 / 5.6	0.7 / 18.1	3.1 / 28.8	3 / 30.5
en_ga	34.4 / 60.4	32.8 / 59.7	33.2 / 60	33.4 / 60.2	0.1 / 6.7	0.6 / 9.2	1.7 / 21.4	2 / 22.9
en_gd	21.6 / 53.6	19.7 / 48	23.2 / 51.8	21.3 / 50.2	0.1 / 2.3	0.2 / 6	0.3 / 14.9	0.2 / 12.8
en_gl	36 / 61.9	37.1 / 63.1	37.1 / 63	37.4 / 63.5	0.2 / 8	29.8 / 57.4	31.5 / 58.8	31.5 / 58.9
en_gn	9.8 / 40.4	11.6 / 40.9	10.9 / 38.9	12.5 / 41.8	0.1 / 3.9	0.1 / 5.6	0.3 / 15.2	0.1 / 9
en_gu	25 / 56.6	18.7 / 50.3	20.1 / 51.4	20.4 / 52.1	0 / 0.2	0.2 / 0.8	0.1 / 2	0 / 1.4
en_ha	28.8 / 56	16.9 / 44.4	18.9 / 45.3	18.7 / 47.1	0.1 / 2.6	0.2 / 6.2	0.2 / 12	0.3 / 13.9
en_hi	34.6 / 59.3	30.4 / 54.1	31 / 54.6	32.2 / 55.8	0.2 / 4.1	17.9 / 46.5	24.7 / 49.7	22.9 / 49.7
en_hne	27.1 / 57.5	19.9 / 51.2	17.8 / 49	20.4 / 52.1	0 / 0.2	1.7 / 23.8	2.4 / 31.2	2.9 / 33.2
en_hr	31.9 / 60.1	32.3 / 61.5	32.1 / 61	32.8 / 61.6	0.9 / 12.5	24.7 / 54.4	25.2 / 55.9	25.9 / 56.8
en_ht	25.1 / 53.9	18.8 / 47.7	23.5 / 52.1	22.5 / 51.3	0.1 / 4.7	0.9 / 15.9	4.1 / 30.2	3.3 / 28.8
en_hu	27.6 / 58.7	29.1 / 60	28.9 / 59.9	29.4 / 60.4	0.2 / 6.3	19 / 49.8	18.8 / 50.9	20.2 / 52.1
en_hy	21.4 / 57.8	15.9 / 47.2	19.5 / 53.7	20.1 / 53.9	0 / 0.2	0.5 / 0.9	0.5 / 0.7	0.5 / 0.7
en_id	46.6 / 70.6	43.5 / 68.5	44.5 / 69	44.7 / 69.5	0.1 / 5.8	34.1 / 62.4	38.6 / 65.5	39 / 65.4
en_ig	16.9 / 43.6	13.1 / 34.6	16.1 / 38.7	16.2 / 38.5	0.1 / 3.4	0.2 / 8.6	0.2 / 8.7	0.1 / 6.5
en_ilo	25.6 / 55.8	15.6 / 42.4	17.7 / 45.7	17.9 / 45.8	0.1 / 4.5	3.7 / 29.6	3.5 / 29.6	1.9 / 28.3
en_is	25.3 / 52.8	25.9 / 53.5	27.5 / 54.2	27.3 / 54.5	0.8 / 11.7	24.1 / 51	24.3 / 51.7	24 / 51.6
en_it	31.8 / 59.9	32.4 / 60.3	32.2 / 60.3	32.7 / 60.5	0.4 / 10.8	24.5 / 54.2	24.8 / 53.7	25.1 / 54.6
en_he	34.7 / 62.6	30.5 / 59.3	32.5 / 61.1	31.4 / 60.6	0.1 / 0.9	15.7 / 46.7	19.9 / 50.6	19.7 / 50.9
en_ja	0 / 35	8.8 / 34.3	9.3 / 35.5	9.4 / 35.9	0 / 1.6	0.2 / 29.3	0 / 30.5	0 / 30.7
en_jv	28.6 / 57.2	12.1 / 37.9	14.5 / 42.2	16.4 / 44.4	0.1 / 5.1	2.7 / 25.4	8.1 / 39.1	9.3 / 40
en_ka	15 / 52.6	2.8 / 17.4	2.2 / 14.8	2.3 / 15.8	0.2 / 5.7	10.9 / 46.4	10.7 / 48.7	10 / 48.6
en_kab	10 / 37.9	0.7 / 1.4	0.9 / 12.4	0.7 / 1.9	0 / 3.4	0.1 / 9.6	0 / 9.6	0 / 5.9
en_kac	12 / 39.1	2.9 / 22.8	3.6 / 23.9	4 / 25.3	0.1 / 3.6	0.3 / 16.1	0.1 / 8.8	0 / 7.6
en_kam	4.3 / 28.8	1.2 / 2.3	2 / 17.7	1.4 / 3.6	0.1 / 4.5	0.7 / 10.3	0 / 4.9	0.1 / 5.7
en_kbp	6.9 / 30.4	1.4 / 14.7	2.3 / 16.4	2.1 / 16.9	0.1 / 2.3	0.1 / 5	0.1 / 5.5	0 / 3.3
en_kea	18.2 / 44.9	1.5 / 3.1	1.8 / 10	1.5 / 3.1	0.5 / 8.7	1.1 / 23.8	0.5 / 19	2.1 / 28.6
en_kg	16.6 / 48.2	0.8 / 14.7	1 / 16.3	1.4 / 19	0.3 / 7.8	0.1 / 7.9	0.1 / 6.9	0 / 6.3
en_ki	11.6 / 40.1	1.2 / 2.2	2.5 / 15.5	1.2 / 1.8	0.4 / 7.3	0.3 / 8.3	0.1 / 6.5	0.1 / 6
en_kk	22.2 / 56	17.1 / 50.2	18.7 / 52.2	18.8 / 53.1	0.5 / 7.7	13.4 / 48.7	12.6 / 50	10.4 / 49.7

en_km	3.7 / 45	5.3 / 40.6	5.2 / 40.7	5.4 / 41.4	0.1 / 0.7	0.9 / 29.2	0.2 / 28.5	0.1 / 24.7
en_kmb	2.9 / 28.3	1.7 / 19.8	1.8 / 19.1	1.4 / 19.4	0.2 / 6	0.2 / 8	0 / 8.4	0 / 4.9
en_kn	21.3 / 58.2	11.9 / 49.2	13.8 / 50.9	13.3 / 51.3	0.1 / 0.4	0.1 / 1.6	0 / 0.8	0 / 0.4
en_ko	14.3 / 38	14.7 / 36.2	14.9 / 36.7	15.1 / 37.1	0.3 / 3.2	5.5 / 24.3	4.9 / 25.5	4 / 25.6
en_kr_Arab	0.3 / 12.7	0.5 / 20.8	0.5 / 20.5	0.5 / 22.2	0.1 / 0.5	0 / 4	0 / 4.1	0 / 2.1
en_kr	3.9 / 31.5	3.8 / 28.4	3.3 / 27.2	4 / 29.9	0.3 / 6.3	0 / 11.4	0.1 / 13.3	0 / 4.4
en_ks_Deva	1.8 / 20.4	1.9 / 21.2	1.6 / 19.8	1.9 / 21.6	0.2 / 1.2	0.2 / 8.3	0 / 4.3	0 / 3.9
en_ks	6.7 / 38.9	6.6 / 35.6	5.8 / 32.9	7.2 / 36.5	0 / 0.3	0.1 / 4.7	0 / 3.6	0 / 2.9
en_ku	12.2 / 42.3	0.3 / 0.5	0.3 / 0.7	0.5 / 0.8	0.2 / 6.4	0.5 / 14	0.4 / 15.4	0.4 / 15.8
en_ky	13.7 / 49.1	10.9 / 44.7	13.4 / 48.1	12.3 / 46.9	0.2 / 1.7	1.5 / 25.9	1 / 27.8	0.6 / 26
en_lb	27.8 / 59.1	14.7 / 38	16 / 38.2	16.9 / 41.3	0.4 / 10.6	3.3 / 34.1	3.5 / 34.9	4 / 35.7
en_lg	9.9 / 44.7	3.5 / 27	4 / 30.3	3.7 / 29.6	0.3 / 7.4	0.4 / 13.9	0.2 / 11.5	0.1 / 10.9
en_li	19 / 51.6	18.4 / 51.1	15.6 / 47.1	18.4 / 51.5	0.8 / 12.5	2.9 / 30.5	5.9 / 38.9	6.5 / 39.8
en_lij	28.7 / 56.1	29.5 / 55.2	24.9 / 50.3	29.5 / 55.4	0.5 / 10.6	1.9 / 25.6	3.5 / 32.1	2.4 / 32
en_lmo	8 / 38.7	6.7 / 32.2	5 / 27.9	6.7 / 32.9	0.4 / 9.1	1.2 / 19.6	1 / 21.8	2.3 / 26.9
en_ln	19.1 / 50.8	2.1 / 18.1	2.3 / 18.8	2.7 / 20.8	0.4 / 9.4	0.7 / 11.4	0.1 / 9.3	0.2 / 10.6
en_lo	6.8 / 53.4	9.9 / 47.1	9.8 / 46.8	9.7 / 50.5	0.2 / 1	1.9 / 35.2	1.4 / 38	1.3 / 39.5
en_lt	27.3 / 58.1	29 / 60.1	29.1 / 60.5	30.1 / 61.1	1.5 / 13.8	22 / 54.3	21.9 / 54.2	22.5 / 54.4
en_ltg	26 / 56.8	25.3 / 57.4	21.5 / 53.3	26.3 / 58.9	0.3 / 8.7	3.2 / 29.9	2.5 / 29.5	3.5 / 31.2
en_lua	6.4 / 39.4	1.4 / 2.2	2.1 / 12	1.5 / 2.6	0.3 / 5.7	0.3 / 12.6	0.3 / 11.6	0.1 / 8.1
en_luo	11.7 / 41.7	1.1 / 2.7	2.4 / 18.3	1.1 / 2.7	0.2 / 2.9	0.1 / 20.8	0.1 / 11.6	0 / 5.4
en_lus	11.3 / 40.9	9.5 / 33.7	11.7 / 37.4	11.3 / 36.7	0.4 / 7.4	0.5 / 11.5	0.2 / 11	0.1 / 7.4
en_lv	28.6 / 57.9	32.8 / 61.8	32.7 / 61.9	32.9 / 62	1.4 / 11.8	25.2 / 54.9	26.1 / 56.1	26.2 / 56.4
en_mag	32.2 / 61.2	26.2 / 55.3	22.5 / 51.6	26.8 / 56.3	0 / 0.3	11.6 / 41.4	6.7 / 41	6 / 40
en_mai	16.6 / 50.7	8.3 / 42.5	12.2 / 45.3	10.3 / 44.3	0.1 / 0.9	4.8 / 35.9	3.3 / 34.7	3.4 / 35.6
en_mg	17.7 / 54.2	17 / 50.8	18.5 / 53.1	18 / 52.8	0.3 / 7.1	1.7 / 20.1	0.5 / 17	0.2 / 13.5
en_mi	18.9 / 45.9	18.6 / 44.1	16.8 / 42	19.1 / 45.2	0.4 / 7.9	0.5 / 11.5	0.4 / 12.1	0.3 / 10.8
#N/A en_min	23.4 / 55.6	12.4 / 44.9	14.1 / 45.6	10 / 43.2	0.2 / 6.5	7.6 / 38.5	8 / 39.2	5.6 / 37.7
en_mk	35.6 / 63.2	36.6 / 64.8	37.7 / 65.2	37.4 / 65	1 / 3.4	30.8 / 59.6	30.5 / 59.8	31.2 / 60.3
en_ml	17.1 / 57.3	9.2 / 47.1	9.4 / 47	10.2 / 49.2	0.5 / 8.6	5.8 / 42.8	6.4 / 46.1	5.6 / 47.7
en_mn	14.8 / 47.7	12.2 / 44.4	14.8 / 47.8	14.7 / 47.7	0.1 / 0.7	0.6 / 0.8	0.6 / 6	0.3 / 2.4
en_mni	7.8 / 44.3	6.6 / 38	7 / 39.2	7.4 / 40.4	0.1 / 0.2	0 / 4.3	0 / 4.8	0 / 3
en_mos	4.5 / 26.4	0.9 / 2.3	1.2 / 16.6	0.8 / 2.2	0.2 / 6.1	0.3 / 14.2	0 / 6.5	0 / 3.5
en_mr	17.6 / 52.1	10 / 39.1	9.9 / 37.7	11.8 / 42	0.8 / 11.9	10.7 / 42.8	8.6 / 43.1	8.6 / 44
en_ms	42.2 / 68.5	37.3 / 64.8	39.3 / 66.2	38.1 / 65.7	1.3 / 13.9	35.8 / 65	37.2 / 65.6	36.2 / 65.4
en_mt	35.6 / 70.8	51.7 / 71.7	52.7 / 72.2	52.5 / 72.2	0.4 / 8.9	5.2 / 34.3	5.4 / 38.5	2.9 / 36.7
en_my	2.9 / 40.1	3.3 / 40.5	3.1 / 39.1	3.5 / 38.3	0 / 0.2	0.1 / 5.7	0 / 9.9	0.1 / 12.9
en_ne	16.7 / 49.7	10.3 / 41.8	10.4 / 42.1	11.7 / 45.2	0.1 / 0.4	3.3 / 34.8	2.6 / 34.4	2.1 / 34.9
en_nl	27.5 / 57.8	28.6 / 59.3	28.7 / 59.4	28.8 / 59.4	1.6 / 15.9	22.5 / 53.7	22.4 / 54.3	23.2 / 54.6
en_nn	28.1 / 56.1	33.6 / 61.3	34.6 / 61.6	34.3 / 61.8	3 / 15.2	27.6 / 56.1	29.2 / 57.2	29.8 / 57.6
en_no	32.9 / 61.2	34.7 / 62.8	34.4 / 62.4	34.5 / 62.8	2.5 / 15.8	29.1 / 57.5	30.8 / 59.3	31.1 / 59.8
en_nso	23.7 / 53.3	12.8 / 37.7	15.7 / 41.1	14.5 / 40.5	0.3 / 4.5	0.6 / 11.8	0.2 / 10	0.2 / 9.1
en_nus	6.5 / 31.5	4.8 / 29.3	3.9 / 27.7	5 / 30.2	0.2 / 4.3	0 / 2.7	0 / 2.1	0 / 0.7
en_ny	13.6 / 48.7	8.1 / 38.6	11.5 / 44.9	10.6 / 43.5	0.4 / 8.1	0.7 / 16.4	0.4 / 15.7	0.4 / 17.2
en_oc	34 / 61.1	38.1 / 63	39.6 / 64.3	39.7 / 64.7	0.4 / 10.5	6.9 / 38.6	7.8 / 40.1	8.6 / 41.1
en_om	5.4 / 43.3	1.1 / 22.8	1.7 / 29.3	1.8 / 29.1	0.1 / 7.4	0.1 / 10.6	0 / 7.9	0 / 6.3
en_or	15.1 / 50.4	12.4 / 46.7	13.3 / 47.5	13.1 / 47.8	0.1 / 0.2	0.4 / 0.9	0.3 / 0.8	0.3 / 0.8
en_pa	24.5 / 51	20.7 / 45.7	21.4 / 46.5	21.1 / 46.5	0.1 / 0.3	0.9 / 6	0.3 / 1.7	0.2 / 1.8
en_pag	17.3 / 49.6	7.1 / 35.4	10.2 / 38.4	11.1 / 39.2	0.6 / 7.4	0.7 / 20.9	4.3 / 30.7	4.9 / 32.2
en_pap	37.8 / 61.5	15.5 / 46.5	20 / 49.4	16 / 47	0.6 / 11.4	2.7 / 28.5	5.7 / 34.6	6.7 / 36.6

en_pl	22.2 / 52.1	23.3 / 53.7	23.6 / 53.8	23.8 / 53.9	1 / 11.6	16.7 / 46.7	16.2 / 47.9	17.5 / 48.3
en_ps	15.5 / 40.8	15.1 / 38.7	15.3 / 38.8	15.4 / 39.1	0.1 / 0.9	0.4 / 12.1	0.2 / 10	0.1 / 8.9
en_pt	47.9 / 69.6	49.9 / 71.5	51 / 72	51.4 / 72.3	3 / 17.8	38.8 / 63.5	41 / 65.1	40.7 / 65.4
en_quy	3.6 / 31.2	1.5 / 22.8	2 / 18.6	1.3 / 25.5	0.2 / 8.7	0.3 / 14.5	0.1 / 8.8	0 / 6.7
en_rn	12.9 / 47.3	5 / 31.8	5.9 / 32.6	5.9 / 30.7	0.3 / 5.5	0.5 / 13.9	0.5 / 15.5	0.2 / 12.9
en_ro	38.4 / 63.6	42 / 66.2	42.3 / 66.2	42.4 / 66.5	4.4 / 20	32.6 / 58.1	34.8 / 60.5	35.4 / 60.9
en_ru	32.3 / 59	32.2 / 59.6	33 / 60.1	32.5 / 59.8	1.1 / 4.4	22.5 / 49.4	23.9 / 52.4	25.4 / 53.2
en_rw	21.1 / 53.8	1.8 / 14.6	0.3 / 9.6	0.9 / 11.9	0.5 / 6.9	0.8 / 15.7	0.5 / 15.5	0.5 / 17
en_sa	1.7 / 31.6	1.6 / 27	1.8 / 30.5	2 / 31.7	0.1 / 0.9	0.1 / 11.6	0.1 / 15.2	0.1 / 14.7
#N/A en_sc	30 / 58.1	24.2 / 52.8	19.3 / 47.4	24.5 / 53.3	0.5 / 10.3	2.7 / 27	0.8 / 21.4	1.1 / 23.8
en_scn	17.7 / 50.5	15.3 / 48.5	13.8 / 45.3	15.6 / 49.1	0.5 / 9.3	1.3 / 22.6	1.5 / 26.3	1.8 / 28
en_sd	23 / 50.1	19.8 / 43.7	17.5 / 42	18.4 / 42.5	0.1 / 0.5	0.8 / 6	0.7 / 13.7	0.4 / 10
en_sg	9 / 38	6 / 30.1	5.8 / 29.2	6.4 / 30.7	0.2 / 8.3	0.2 / 10.1	0.2 / 10.6	0.1 / 5.3
en_shn	5.5 / 42	5.4 / 37.5	5 / 37.3	5.6 / 39.6	0.2 / 1.2	0.1 / 3.1	0 / 2.6	0 / 3.6
en_si	14.7 / 48	13.5 / 46	13.4 / 45.2	14.2 / 47.6	0 / 0.4	0.5 / 1	0 / 0.7	0.1 / 0.8
en_sk	35.2 / 61.6	35.7 / 62.5	36.3 / 62.9	36.8 / 63.1	1.5 / 13.9	27.7 / 55.4	28.6 / 56	26.5 / 55.8
en_sl	31.7 / 58.7	32.3 / 59.9	32.9 / 60.2	33.3 / 60.6	0.8 / 12.7	23.7 / 52.4	25.1 / 54.1	24.8 / 53.3
en_sm	26.8 / 50.9	17.9 / 43.6	20.8 / 46.5	20.5 / 46.8	0.5 / 7.4	0.6 / 14	0.3 / 11.4	0.2 / 8.9
en_sn	12.6 / 48.8	7 / 38.2	9.3 / 42.3	8.8 / 41.8	0.5 / 7	0.8 / 16.6	0.4 / 17	0.3 / 16
en_so	12.6 / 47.3	9.4 / 41.8	10.3 / 43.4	10.2 / 43	0.2 / 6.8	0.7 / 15.3	0.2 / 12.9	0.2 / 13.2
en_sq	33.2 / 60.6	29.8 / 58.6	30.9 / 59.5	30.4 / 59.3	1.2 / 11.7	28.5 / 56.8	28.7 / 56.6	29 / 57.7
en_sr	35.9 / 62.1	34.9 / 61.6	35.8 / 62.4	36.2 / 62.7	0.5 / 2	29.2 / 57.4	29.9 / 58.4	30.4 / 58.1
en_ss	10.7 / 49.2	4.1 / 33.4	5.1 / 34.7	5.3 / 36	0.5 / 7.2	0.5 / 14.9	0.1 / 11.5	0.1 / 9.5
en_st	18.9 / 49.2	9.6 / 34.7	15.4 / 43.1	14.1 / 41.6	0.6 / 8.7	1 / 15.3	0.5 / 13.3	0.3 / 11.9
en_su	15.9 / 48.3	14.4 / 45.5	18.7 / 50.6	18.5 / 50.4	0.5 / 9	5.8 / 35.5	7 / 37.9	7.8 / 38.3
en_sv	44.2 / 68.1	47 / 70.8	47.2 / 70.8	47 / 70.8	2.5 / 15.8	37.7 / 64.2	39.4 / 65	39.1 / 65.3
en_sw	32.6 / 61.2	29.3 / 57.6	30.4 / 58.1	30.4 / 58.5	0.3 / 9	1.5 / 21.1	1.4 / 22.8	1.3 / 25.7
en_szl	27 / 56.7	24.9 / 56.1	21.6 / 51.8	26.2 / 56.7	0.5 / 9.3	6.9 / 36.7	4.9 / 35.1	4.3 / 35.8
en_ta	19.8 / 59.4	12.4 / 51.4	13 / 52.6	13.9 / 54.1	0.4 / 10.7	8.6 / 47.8	7.8 / 49.2	8.1 / 49.5
en_taq_Tfng	0.7 / 19.3	0.6 / 21.2	0.5 / 20.4	0.7 / 21.7	0.1 / 0.4	0 / 7.8	0 / 2	0 / 1.4
en_taq	4.2 / 25.9	3 / 24.6	3.3 / 24.7	3.6 / 25.7	0.2 / 8	0.1 / 4.7	0 / 2.8	0 / 4.7
en_te	24.8 / 60.1	16.7 / 51.4	17.6 / 52.1	17.5 / 52.7	0.7 / 9	10.9 / 45.9	8.3 / 47.5	10.5 / 49.6
en_tg	24 / 54	6.1 / 26.6	9.2 / 31.3	10.2 / 32.6	0.1 / 0.6	1.4 / 21	0.8 / 22.6	0.9 / 24.4
en_th	6 / 51.4	8 / 44.1	7.6 / 43.1	8 / 44.3	0.1 / 10	7.1 / 48.5	7.5 / 48.7	4 / 49.2
en_ti	6 / 27.9	1.8 / 15.7	2.5 / 18.6	2.5 / 18.9	0.1 / 0.3	0 / 1	0 / 0.8	0 / 0.6
en_tk	13.7 / 46.2	13.9 / 44.5	17.5 / 50.7	17.3 / 49.9	0.2 / 6.9	0.8 / 19.9	0.9 / 22.9	0.7 / 22.6
en_tn	23 / 51	10.6 / 34.8	14 / 38.6	12.2 / 37.7	0.5 / 8.2	0.6 / 14.5	0.3 / 11.5	0.2 / 10
en_tpi	18.3 / 42.8	1.1 / 2.7	3.4 / 22.2	1 / 2.3	0.6 / 5.4	0.7 / 16.7	1.4 / 22.8	0.3 / 11.3
en_tr	30.1 / 61.8	29.9 / 60.5	31 / 61.2	30.5 / 61.4	1.2 / 14.4	16.9 / 51	15 / 51.6	15.7 / 52.1
en_ts	23.3 / 53.2	9.1 / 34.7	12 / 38.1	10.3 / 37	0.5 / 7.8	0.3 / 6.9	0.2 / 10.5	0.2 / 10.3
en_tt	18.4 / 50.6	7 / 31.5	10.2 / 36.5	13.1 / 42.4	0.1 / 1	1.8 / 25.1	0.5 / 21.7	0.5 / 23.4
en_tum	10.5 / 38.5	1.4 / 2.2	2.2 / 15.3	1.5 / 2.5	0.4 / 7.4	0.2 / 14.8	0.4 / 13.8	0.1 / 10
en_ug	13.8 / 50.6	5.9 / 33.4	6.9 / 34.2	8.8 / 39.8	0 / 0.3	0.2 / 1.9	0.1 / 1.2	0.1 / 1.6
en_uk	31.5 / 59.1	30.6 / 58.9	30.9 / 59.4	31.1 / 59.4	1.6 / 7.8	23.3 / 52.3	24.4 / 53.5	23.5 / 53.4
en_umb	2.8 / 31.2	1.6 / 2.9	1.4 / 10.5	1.4 / 2.5	0.2 / 6.1	0.1 / 11.5	0 / 5.6	0 / 4.6
en_ur	23.7 / 51.1	18.8 / 44	20.1 / 44.8	20.3 / 45.4	0.1 / 0.8	7.7 / 33.4	6.1 / 34.5	2.9 / 30.3
en_uz	17.6 / 55.7	0.6 / 1.3	0.5 / 1.1	0.6 / 1.1	0.1 / 6.8	2.5 / 30.3	3.5 / 38.5	4.6 / 43.9
en_vec	22.8 / 54.8	25.1 / 56.4	22 / 52.2	25.2 / 56.7	0.4 / 12.8	3.5 / 29.6	5.6 / 37.7	7.6 / 41.2
en_vi	42.3 / 59.6	39.1 / 57.2	40.5 / 58.2	40.5 / 58.6	2.2 / 10.3	35.2 / 54	35.7 / 54.7	34 / 54.4
en_war	31.7 / 59.3	9.8 / 31.1	18.5 / 44.9	13.7 / 44	0.7 / 7.5	5.9 / 33.5	8.7 / 38.8	10.4 / 40.3

en_wo	6.8 / 31.8	1.2 / 13.2	1 / 12.6	1.4 / 15.4	0.3 / 7.7	0.1 / 5.5	0.1 / 7	0 / 3.5
en_xh	15.2 / 54.7	11.3 / 47.4	12.5 / 50.1	12.5 / 49.2	0.2 / 7.7	0.4 / 15.4	0.3 / 16.1	0.2 / 16.2
en_yi	11.9 / 41.8	8.7 / 34.8	6.6 / 36.3	3.9 / 32.5	0.2 / 0.7	0.8 / 13.6	0.9 / 17.6	0.5 / 16.3
en_yo	6 / 27.4	2.6 / 19.7	2.7 / 20.3	2.9 / 20.7	0.3 / 6.3	0.1 / 10.6	0.1 / 5.2	0 / 3.2
en_yue	2 / 20.8	0.3 / 2.2	0.2 / 2	0.3 / 2	0.1 / 2.5	0.1 / 24.7	1.2 / 25.5	0.8 / 25.3
en_zh_Hant	15.4 / 17.6	5.5 / 30.4	6.3 / 31.8	36.5 / 32.1	0.1 / 3	1.2 / 21.4	3.6 / 22.3	1.3 / 21.9
en_zh	31.2 / 29.3	5.8 / 35.7	6.1 / 36.6	41.3 / 36.5	2.9 / 3.3	29.2 / 26.2	30.1 / 26.7	29.8 / 26
en_zu	19.7 / 58.8	11.2 / 44.7	13.8 / 48.6	13.4 / 47.6	0.3 / 7.1	0.4 / 13.9	0.3 / 17	0.2 / 17
ace_Arab_en	13.6 / 37.9	17.1 / 43.5	16.6 / 42.7	19.1 / 45.8	0 / 0.7	0.8 / 13.6	1.5 / 16.5	1.6 / 18.5
ace_en	31.3 / 53.9	23.6 / 49	21.1 / 46.9	25 / 50.8	0 / 3	3.3 / 22.9	8.5 / 32.8	6.8 / 33.7
acm_en	40.1 / 65	34.2 / 60.8	35.8 / 61.7	35.2 / 61.6	0.1 / 1.6	25.9 / 52.3	28.4 / 55.3	28 / 55.1
acq_en	42.4 / 66.7	36.5 / 62.5	37.3 / 62.9	37.4 / 63.3	0.1 / 2	23.5 / 51.4	30.1 / 57.1	29.5 / 57.1
aeb_en	36 / 61.4	28.5 / 55.5	30.4 / 57.1	30.1 / 56.9	0.1 / 2.4	19 / 44.5	22.4 / 49	24.3 / 51.3
af_en	59.2 / 77.1	58.7 / 76.8	60.1 / 77.9	60 / 77.7	2.4 / 18	51.7 / 71.5	54.3 / 74	54.9 / 74
ajp_en	46.6 / 68.9	38.1 / 62.8	37.7 / 62.4	39.6 / 64	0.1 / 1.5	32.7 / 58.3	32.7 / 58	31.2 / 58.7
am_en	36.5 / 61.8	32.6 / 57.9	36.9 / 61.1	35.9 / 60.6	0.2 / 4.8	20.3 / 47.1	22.7 / 49.6	22.9 / 50
apc_en	42.5 / 66.7	34.1 / 60.6	34.7 / 60.7	36.2 / 62.1	0.1 / 2	26.2 / 52.1	28.7 / 55	29.8 / 56.3
ar_MA_en	32.2 / 0	23 / 50.3	24.5 / 51.2	25.1 / 51.7	0 / 1.4	15.3 / 43.4	19.3 / 45.5	18.1 / 46.8
ar_en	45.3 / 68.6	41.7 / 66.4	44.4 / 68.1	43.1 / 67.5	0.3 / 4.9	32.9 / 58.1	34.9 / 60.7	35.5 / 60.6
ars_en	43.8 / 67.8	40.3 / 65.4	42.7 / 66.7	41.7 / 66.5	0.1 / 2	28.3 / 52.4	34.7 / 59.9	34.5 / 60
arz_en	37.2 / 62.7	30.2 / 57.5	29.1 / 56.5	31.7 / 58.9	0.1 / 1.5	24.7 / 51.1	25 / 52.9	25.3 / 53.4
as_en	33.9 / 59.8	29.5 / 55.7	31.8 / 57.4	31.2 / 57.3	0.1 / 1.6	14.7 / 43.3	19.4 / 47.4	19.4 / 47.9
ast_en	43 / 66.4	37.3 / 62.5	39.7 / 64.2	39.8 / 64.5	1.2 / 13.3	26.2 / 56.1	29.2 / 57	31.6 / 58.3
awa_en	42.8 / 67.4	34.6 / 61.6	33.5 / 60.2	36.6 / 62.9	0.1 / 2.5	19.8 / 47.2	23.9 / 53.4	23.8 / 53.8
ay_en	12.1 / 33.7	6.1 / 27.4	7.9 / 28.4	8.3 / 29.6	0 / 2.3	0.5 / 16.4	3 / 18.8	2.8 / 19.1
az_en	27.2 / 56.5	19.7 / 47.7	20.7 / 47.2	25.1 / 52.7	1.3 / 15.7	18.8 / 47.5	20.2 / 49.5	20 / 48.7
azb_en	20.8 / 47.7	4.8 / 30.2	6.7 / 31.7	6.3 / 32	0.1 / 3.4	2.1 / 23.3	4.2 / 31.2	5.4 / 30.1
ba_en	33.6 / 60.1	6.7 / 25.7	8.2 / 26.8	10 / 29.2	0.1 / 2.9	13.1 / 36.1	19.7 / 46.7	19.3 / 46.8
ban_en	38.2 / 62.3	29 / 55.6	25.5 / 52.4	30.3 / 57	0.2 / 5.4	8.8 / 34.8	15.9 / 43.9	16.7 / 43.9
be_en	23.7 / 55.4	17.9 / 49.3	17.6 / 49.4	21.6 / 52.9	0.8 / 11.5	18.4 / 48.3	19.6 / 50.9	19.6 / 51.5
bem_en	29.6 / 52.4	7.4 / 30.4	8.6 / 31.5	10 / 33.2	0 / 2.5	1.2 / 14.7	5.4 / 22.2	4 / 21.9
ber_en	19.7 / 0	16.6 / 41.6	14.6 / 39.9	17.6 / 43.1	0 / 1.7	0.1 / 14.8	0.5 / 13.7	0.2 / 11.9
bg_en	43.1 / 68.3	42.2 / 67.8	43.6 / 68.9	43.2 / 68.5	1.8 / 15.5	34.1 / 62.7	36.8 / 64.1	37.8 / 64.5
bho_en	34.6 / 60.7	28.5 / 56	25.5 / 53.5	28.9 / 56.8	0.1 / 2.4	16.5 / 43.8	19.6 / 48.5	17 / 48
bjn_Arab_en	19.2 / 43.3	21.7 / 47.2	20.3 / 46.1	23.8 / 49.8	0 / 0.9	0.5 / 15.3	1.7 / 18.1	1.8 / 18.1
bjn_en	40.1 / 63.3	30.2 / 55.9	26.1 / 52.3	32 / 57.5	0.1 / 3.5	10.9 / 40.4	17.6 / 44.4	20.3 / 45.5
bm_en	19.6 / 42.2	14.1 / 37.1	12.9 / 36.3	15.5 / 39.4	0 / 2	1.7 / 12.8	1.9 / 17.8	2.3 / 18.8
bn_en	38.7 / 64.2	34.9 / 60.6	37 / 62.5	36.2 / 61.8	0.2 / 2.7	21.8 / 50.7	25.2 / 53.9	26.3 / 54.8
bo_en	15.6 / 40.9	12.6 / 38.1	12.6 / 38.2	14.6 / 41.1	0 / 1	1.9 / 20.5	5.9 / 29.6	5.6 / 29
bs_en	45.2 / 68.8	44.1 / 68.5	45 / 69.1	44.9 / 69.2	1.5 / 13.7	36.2 / 62.3	38.5 / 64.5	39 / 65
bug_en	23.1 / 48.4	17.7 / 43.6	15.7 / 42	18.9 / 45.3	0.1 / 3.3	1.8 / 18.7	5.6 / 28	5.9 / 27.6
ca_en	49 / 71.1	48.5 / 71	50.3 / 72.2	50 / 72	2.2 / 16.1	41.8 / 66.6	42.5 / 67.1	43 / 67.2
ceb_en	45.6 / 66.8	34.8 / 60.2	34.1 / 59.4	36.6 / 61.2	0.3 / 5.7	31.9 / 56.4	32.2 / 57.2	33.6 / 58.5
cjk_en	11.3 / 32.7	4.8 / 23.8	5 / 24.7	5.6 / 26.5	0 / 2.3	1.2 / 16.2	2 / 17.7	2.3 / 18.8
ckb_en	37.5 / 61.6	28.3 / 53.1	34.4 / 59	34.2 / 59	0.2 / 4.4	18 / 42.1	22.5 / 48.6	22.8 / 49.5
crh_Latn_en	35.9 / 61.8	29.2 / 56.8	25 / 53.1	29.1 / 57	0.1 / 4.8	16.1 / 40.3	18.8 / 45.2	17.8 / 46.6
cs_en	41.2 / 66.4	41.3 / 66.6	42.3 / 67.3	42.3 / 67.3	2.1 / 18.1	34.7 / 61.4	35.8 / 62.5	36.1 / 62.9
cy_en	60 / 77.1	57.5 / 75.6	60.6 / 77.7	60.8 / 77.7	1.2 / 12.5	43.9 / 65.5	48.5 / 69.2	47.5 / 68.6
da_en	49.5 / 71.2	50.1 / 72.3	51.8 / 73.3	50.8 / 72.8	1.6 / 15.1	44.5 / 68.2	44.1 / 68.9	44.7 / 68.9
de_en	45.8 / 69.2	45.5 / 69.5	46.4 / 70.1	46.5 / 70	1.6 / 15.9	37.8 / 63.3	39.5 / 64.9	38 / 64.6

din_en	12.7 / 32.9	10.3 / 31.1	9.5 / 31.1	11 / 32.8	0 / 2.2	1.2 / 18.8	1.6 / 16.3	2.4 / 19.3
dyu_en	8.5 / 30	7.9 / 31.5	7.9 / 31.3	8.9 / 33.5	0 / 2.3	1.1 / 15.3	1.9 / 17.4	2.3 / 18.2
dz_en	17 / 44	12.9 / 40.4	12 / 39.2	13.7 / 41.6	0 / 0.8	0.7 / 16.4	2 / 22.4	1.9 / 22.7
ee_en	17.8 / 42	10.3 / 33.3	14.3 / 37.4	13.3 / 36.6	0.1 / 4.7	2.5 / 17.5	5 / 23.2	4.4 / 24.3
el_en	40 / 64.9	37.8 / 63.3	38.9 / 64.2	38.8 / 64	1.3 / 12.8	32.3 / 58.4	32.9 / 59.7	33.1 / 60
eo_en	47.8 / 70.1	45.8 / 68.5	47.2 / 69.4	47.1 / 69.5	9.8 / 31	38.3 / 63.3	40.4 / 64.7	40.9 / 65.3
es_en	33.8 / 61.4	31.5 / 61	32.2 / 61.3	32.5 / 61.4	4.5 / 23.9	26.4 / 56	27.6 / 57.8	27.7 / 57.9
et_en	38.8 / 64.6	39.4 / 64.9	40.3 / 65.6	39.6 / 65.5	7.9 / 28	31.8 / 59.2	32.5 / 59.8	32.6 / 59.8
eu_en	34.4 / 61.2	32.6 / 58.6	34.3 / 60	34.4 / 60.4	4.4 / 23	25.5 / 53.8	24.7 / 52.9	26.2 / 54.5
fa_AF_en	41.3 / 0	37.9 / 63	38.7 / 63.2	39.4 / 64.1	2.2 / 11.6	28.7 / 53.7	31.1 / 56.4	32.4 / 57.8
fa_en	40.9 / 65.6	37.7 / 62.9	40.3 / 64.8	39.4 / 64.3	3.9 / 16.4	31.1 / 57.6	30.8 / 57.2	32.3 / 58.7
ff_en	13.5 / 0	10.5 / 33.2	10.1 / 33.1	12.4 / 35.8	0.2 / 10.5	2.6 / 15.7	2.5 / 16.9	2.4 / 18.2
fi_en	36.9 / 63	36.1 / 62.6	36.7 / 63	37.1 / 63.3	3.8 / 22.9	27.9 / 55.8	29.8 / 57.4	30.5 / 58.3
fil_en	51.5 / 71.3	47.1 / 68.2	49.5 / 70	49.6 / 70.1	2.5 / 17.9	40.4 / 63.9	40.7 / 64.4	39.7 / 64.6
fj_en	21.6 / 45.4	12.2 / 36	14.6 / 39.2	14.5 / 38.8	0.3 / 10.8	5.1 / 24.8	5.2 / 24.1	4.3 / 26.2
fo_en	37.1 / 59.5	27.5 / 52.8	31 / 56.1	36.7 / 59.9	6.1 / 26	30.1 / 54.4	32.7 / 56.5	33.9 / 57.4
fon_en	12.9 / 35.4	3.9 / 25.1	5.6 / 26.3	6.4 / 27.8	0.1 / 7.9	0.8 / 12.6	1.5 / 17.8	1.8 / 18.3
fr_en	47.9 / 70.1	46.1 / 69.5	47.2 / 70.3	47.1 / 70.1	2.8 / 21.5	39.1 / 64.5	39.9 / 64.9	40 / 65.1
fur_en	45.3 / 68.6	36.9 / 63	30.8 / 58	37.4 / 63.4	3.6 / 23.8	24.2 / 50.3	24.2 / 53.4	23 / 54.7
ga_en	45.9 / 68.1	43.4 / 67	45.8 / 69	45.9 / 68.8	3.3 / 21.1	29.4 / 53.8	31.9 / 57.4	34.2 / 59.4
gd_en	36.8 / 60.9	33.4 / 58.9	37.3 / 61.9	37.5 / 61.6	2.1 / 18.1	21.7 / 46.8	22.8 / 49.1	22.1 / 49.6
gl_en	45.6 / 68.9	44.1 / 68.5	46 / 69.7	45.8 / 69.6	5.3 / 19.6	35.9 / 62.7	36.6 / 63.8	36.9 / 64.4
gn_en	27.7 / 52	19.5 / 45	18.2 / 44	20.8 / 46.9	0.6 / 11.9	5 / 26.3	8 / 30.8	8.1 / 29.9
gu_en	44.6 / 68.4	39.7 / 64.9	41.9 / 66.1	41.2 / 65.5	2.1 / 13.4	25.4 / 54.1	28 / 56.7	28.7 / 57.2
ha_en	36.4 / 58.7	31.8 / 54.5	35.9 / 57.9	35.4 / 57.6	0.8 / 15	17.2 / 40.1	21.6 / 44.8	21.1 / 45.7
hi_en	44.4 / 68.3	40.4 / 65.1	42.9 / 66.6	41.7 / 65.9	2.9 / 12.9	26.1 / 53.9	30.9 / 58.8	32.8 / 60.1
hne_en	52.2 / 73.8	40.1 / 65.5	35.1 / 61.7	41.9 / 66.9	0.5 / 4.9	22.4 / 52.1	21.5 / 53.8	23.2 / 54.5
hr_en	38.7 / 64.4	39 / 64.8	39.9 / 65.4	39.6 / 65.3	6.4 / 25.6	33.8 / 60.3	34.9 / 61	35.8 / 62
ht_en	39.9 / 64	39.3 / 63.5	40.5 / 64.3	40.9 / 65.1	0.9 / 10.9	26.9 / 52.6	29.4 / 55	29.4 / 56.5
hu_en	37.3 / 63.7	37.5 / 63.9	38.2 / 64.4	38.1 / 64.5	3.6 / 22.9	29.1 / 57.1	29.6 / 58	29.7 / 58
hy_en	42.3 / 67	33.5 / 59.4	40.7 / 65.2	40.1 / 64.8	2.5 / 13.6	27 / 54	31.1 / 58.5	31.6 / 59.7
id_en	46.5 / 69	42.6 / 66.6	45.3 / 68.4	44.1 / 67.7	2.4 / 16.1	38.4 / 63.4	38.8 / 63.6	37.7 / 63.1
ig_en	33.2 / 55.9	25 / 48.5	29.7 / 52.8	29.2 / 52.7	0.5 / 11.9	10.8 / 31.5	11.3 / 32.3	9.6 / 36.1
ilo_en	41 / 63.6	24.2 / 50.6	33.3 / 57.3	32.5 / 57	0.6 / 12.6	19.8 / 44.5	19.8 / 44	20.6 / 47.2
is_en	34.9 / 59.3	36.9 / 61	38 / 62.3	37.7 / 62	7.2 / 28.5	31.8 / 57.2	33 / 58.7	32 / 57.8
it_en	36.6 / 63.3	34.7 / 63	35.4 / 63.6	35.6 / 63.6	4.7 / 24	29.9 / 57.9	29.4 / 59.5	29.8 / 59.7
he_en	46.5 / 69	43.3 / 65.9	46.4 / 68.4	45.5 / 68.1	4.1 / 13.9	38.3 / 63	39.7 / 64	38.9 / 63.9
ja_en	30.3 / 58.2	27.6 / 55.7	28.8 / 55.9	28.9 / 56.9	0.1 / 0.6	19.3 / 49.8	19.6 / 48.8	20.3 / 49.2
jv_en	43.6 / 65.2	25.9 / 51.9	29.9 / 55.4	29.7 / 54.8	1.7 / 14.5	27.4 / 52.3	28.3 / 52.8	29.2 / 53.6
ka_en	31.4 / 59.5	21.5 / 48.9	21.3 / 48.8	28.7 / 55.6	1.9 / 9.9	22.4 / 50.9	24.4 / 53.6	25.2 / 54.7
kab_en	27.5 / 50.2	6.6 / 28.2	8.2 / 30.6	8.3 / 31.1	0.3 / 10.4	1.8 / 16.8	3.9 / 19.3	2.9 / 19.5
kac_en	18.6 / 43.7	6.3 / 30.1	9.2 / 32.4	10.1 / 34.1	0.2 / 8.2	1.4 / 13	3.2 / 20	2.9 / 20.2
kam_en	13.7 / 35.7	6.6 / 26.8	7.4 / 27.5	8.8 / 29.6	0.2 / 8	2.3 / 18.1	2.9 / 20.5	3.4 / 20.5
kbp_en	13.1 / 35.7	7.9 / 30.8	9.5 / 32.6	9.7 / 32.8	0.2 / 8.6	1 / 9.9	2.3 / 20.6	2.3 / 21.3
kea_en	48.1 / 70.1	34.5 / 60.5	33.6 / 59.1	36 / 61.7	0.6 / 10	16.8 / 41.2	23.3 / 49	23.6 / 50.5
kg_en	21.1 / 44.6	6.9 / 30.6	8.7 / 32.2	9.3 / 33.3	0.1 / 5.2	2 / 18.1	3.4 / 20.1	2.8 / 21.5
ki_en	25.6 / 49.5	7.6 / 30.6	8.7 / 30.6	9.5 / 32.5	0.1 / 8	2.5 / 18.4	3.3 / 19.3	2.6 / 20.7
kk_en	36 / 61.9	32.2 / 58.7	33.8 / 59.4	33.6 / 59.6	3.6 / 14.3	24.8 / 52.2	26.9 / 54.4	26.3 / 54.1
km_en	35.6 / 60.6	33.7 / 59.6	36.5 / 61.3	35.6 / 60.9	1.5 / 13.1	24.5 / 52.4	25.4 / 53.3	26.8 / 54.5
kmb_en	13.7 / 35.7	5.7 / 27	6.5 / 27.6	7.9 / 29.3	0.1 / 8.9	2.2 / 15.3	2.9 / 18.5	2 / 18.3

kn_en	36.9 / 63.1	31 / 58.3	32.8 / 59	32.6 / 59.4	0.7 / 6.8	17.6 / 45.5	20.9 / 50.6	22.9 / 52.3
ko_en	31.2 / 59	30.8 / 58.6	33.1 / 60.2	32.5 / 59.9	0.7 / 2.9	20.6 / 49	23.4 / 52	22 / 52.3
kr_Arab_en	3.2 / 21.2	5.4 / 24.1	5.2 / 24.1	6.9 / 26.2	0.1 / 5.5	0.6 / 12.1	1.2 / 16	0.8 / 14.5
kr_en	16.1 / 38.7	13.1 / 36.9	12.3 / 36.5	14.4 / 39	0.2 / 11.6	1.6 / 17.4	3.5 / 18.9	2.9 / 19.5
ks_Deva_en	27 / 52.7	21.5 / 48.2	19.8 / 46.5	23.5 / 50.4	0.3 / 5.3	4.2 / 27	3.5 / 26.3	4.7 / 27.7
ks_en	36.7 / 62.2	26.4 / 54	25.8 / 53.4	29.8 / 56.6	0.4 / 6.6	5.4 / 28.9	7.3 / 30.4	4.2 / 28.9
ku_en	29.8 / 54.6	29.1 / 54.2	30.8 / 55.5	32.2 / 56.8	1.8 / 16	16.7 / 40.9	21.6 / 47.4	21.6 / 49.1
ky_en	24.1 / 52.5	20.2 / 47.4	26.6 / 54.4	24.8 / 52.4	1.7 / 12.8	14.2 / 41.2	17.6 / 46.1	16.6 / 45.4
lb_en	49 / 71.2	19.1 / 47.2	21.3 / 49	23.1 / 50.8	3.5 / 20.6	35.3 / 60.6	38.2 / 63	36.8 / 63.2
lg_en	25.3 / 48.2	13.2 / 34.9	16.5 / 38.7	15.8 / 38.4	0.3 / 10.9	5.7 / 23.2	5.4 / 25.5	6.1 / 27.2
li_en	44.6 / 67.1	35.4 / 60.6	29.5 / 55.9	35.8 / 61.3	4.9 / 27.4	25.9 / 51.4	27.5 / 52.9	28.5 / 53.3
lij_en	49.2 / 70.8	40.6 / 65.5	33.6 / 60.2	40.3 / 65.3	4 / 24.5	26 / 53.5	26.2 / 54.7	24.4 / 55.2
lmo_en	42.9 / 66.3	34.8 / 60.3	30.2 / 56.8	35.7 / 61.3	2.3 / 20.6	20.6 / 47.6	22.2 / 50.2	23 / 50.9
ln_en	28.9 / 52.1	5.7 / 27.4	5.8 / 27.2	6.8 / 29.5	0.3 / 11.3	6.6 / 27.3	6.6 / 25.5	6.8 / 29.9
lo_en	39.5 / 63.4	35.1 / 60.1	37.7 / 61.9	38.4 / 63.2	1.1 / 10.2	26.1 / 52.8	28.2 / 55.3	28.9 / 56.9
lt_en	35.5 / 61.1	35.1 / 61.4	36.8 / 62.2	36.2 / 62.1	4.5 / 21.9	28.9 / 56.4	30.1 / 57.3	30.1 / 57.5
ltg_en	40.4 / 65.3	35.3 / 62	30.3 / 57.8	36.5 / 62.9	4.4 / 26	25.8 / 52.4	25.6 / 52.1	26.6 / 53.4
lua_en	19.9 / 43.1	6.3 / 27.8	6.7 / 28	7.9 / 30	0.2 / 7.9	4.3 / 21.3	4.2 / 21.9	4.3 / 23.8
luo_en	22 / 45.4	4.1 / 23.6	4.4 / 24.3	5.4 / 26.4	0.2 / 9	1.6 / 19.4	2.8 / 19.1	2.9 / 18.8
lus_en	18.4 / 43.6	13.1 / 37.7	18.5 / 43.2	18.4 / 43.8	1.1 / 15.1	9.8 / 32.1	9.9 / 31.9	10.9 / 34.4
lv_en	37.1 / 62.9	39 / 64.7	40.7 / 65.7	40.1 / 65.5	3.4 / 20	32.9 / 60.1	33.3 / 60.2	33.3 / 60.8
mag_en	51.6 / 73.7	40 / 65.8	35.8 / 62.5	42.1 / 67.3	0.5 / 4.6	24.2 / 53.5	26.6 / 55.7	26.8 / 57.2
mai_en	46.7 / 70.2	35.4 / 62.4	34.1 / 60.9	38.1 / 64.4	1 / 7.4	21.5 / 50.4	23.3 / 52.4	25 / 54.9
mg_en	35.1 / 59.1	30.4 / 55	33.8 / 57.7	34 / 58	1.7 / 18.5	15.6 / 39.5	19.1 / 44.7	20 / 45.9
mi_en	31 / 54.1	20.8 / 45.2	18 / 43.1	21.6 / 46.8	1.9 / 18.7	10.1 / 32	13.4 / 39.1	12.6 / 39.3
#N/A min_en	41.7 / 64	27.3 / 52.9	25 / 51.2	30.6 / 55.9	0.6 / 12.5	19.4 / 43.4	23.4 / 47.7	21.7 / 48.7
mk_en	45.4 / 68.7	44.7 / 68.4	45.9 / 69.1	46 / 69.2	7.7 / 28	37.7 / 62.8	39.9 / 64.9	39.7 / 64.9
ml_en	39.1 / 64.9	33 / 59.4	34.3 / 59.9	35 / 61.1	0.9 / 4.8	19.2 / 47	23.8 / 52.6	23.9 / 53
mn_en	28.9 / 56.3	28.7 / 55.9	32.2 / 58.9	31.9 / 58.8	0.8 / 8	18.5 / 45.5	20.5 / 48	20.1 / 47.9
mni_en	27.8 / 53.5	23.8 / 51.1	22.6 / 49.5	24.1 / 51.2	0.1 / 2.2	1.9 / 20.5	3.3 / 22.3	2.2 / 21
mos_en	12.2 / 34.3	3.4 / 23	3.4 / 22.4	3.7 / 24.1	0.3 / 9.9	1.4 / 15	1.4 / 18	1.6 / 17.5
mr_en	40.3 / 65.8	35.2 / 61.3	36.3 / 61.6	36.9 / 62.4	1.7 / 7.8	24.8 / 53.7	27.8 / 56.3	28 / 56.4
ms_en	47.7 / 69.7	43.9 / 67.1	46 / 68.7	45.7 / 68.6	2.7 / 18.7	38.5 / 63.4	39.9 / 64.6	40.3 / 64.8
mt_en	59.2 / 77.6	57.6 / 77.1	59.2 / 78.1	59.4 / 78.4	4.6 / 22	44.8 / 67.3	46.1 / 68.8	47 / 70.2
my_en	31.5 / 58.4	28.7 / 55.1	30.9 / 56.7	31 / 57.2	0.2 / 4.2	16.1 / 45.6	16 / 43.7	17.7 / 45.9
ne_en	44.5 / 68.8	39.8 / 65.6	41 / 66.2	41.8 / 67.2	1.1 / 7.5	25 / 54.4	28.2 / 56.7	27.1 / 55.7
nl_en	34.4 / 61.2	34.3 / 61.2	34.3 / 61.3	34.3 / 61.4	6.5 / 28.8	27.1 / 55.2	28.8 / 58.1	29.1 / 57.7
nn_en	46.4 / 68.8	47 / 69.7	48.3 / 70.5	47.8 / 70.4	4.8 / 26.9	42 / 65.8	42.3 / 66.5	43.1 / 66.9
no_en	45 / 68	45.1 / 68.4	46.6 / 69.5	45.9 / 69	5.5 / 26.2	40.5 / 65	40.7 / 65.2	39.9 / 65.7
nso_en	42.6 / 63	26.5 / 48.3	31.8 / 53.2	31.9 / 53.2	0.4 / 11.4	9.1 / 26.9	11.3 / 33.8	12.7 / 36.1
nus_en	18.6 / 41.2	12.9 / 36.1	11.9 / 35.2	14.8 / 38.3	0.1 / 6.9	0.5 / 11.9	1 / 16.1	0.6 / 16.3
ny_en	29.2 / 52.9	24.3 / 47.2	27.2 / 50.1	27.2 / 50.5	0.6 / 12	8.4 / 27.7	12.2 / 35.4	13.9 / 37.2
oc_en	58.4 / 77.2	55 / 75.4	58.9 / 77.6	57.4 / 76.8	7.8 / 27.7	42.5 / 66.5	43.8 / 67.3	42.9 / 68.2
om_en	27.2 / 52.1	11.1 / 35.6	17.3 / 42.1	17 / 42.6	0.4 / 10.5	4.1 / 22.4	5 / 27.1	5.9 / 29
or_en	41.6 / 66.2	34.9 / 61	36.1 / 61.5	36.1 / 61.7	1 / 8.6	19.1 / 47.4	20.6 / 49.1	22.5 / 50.7
pa_en	44.8 / 67.9	39 / 63.7	42 / 65.8	41.1 / 65.2	1.5 / 11.5	24.9 / 52.8	26.3 / 53.7	27.2 / 55.2
pag_en	27.9 / 52.3	22 / 46.4	22.9 / 46.2	24.4 / 48.5	0.5 / 11	17.6 / 40.7	17.9 / 42.2	17.7 / 42.8
pap_en	53 / 73.4	48.2 / 70.8	49.7 / 71.7	53.4 / 73.9	1.4 / 15.8	33.1 / 58.1	37.4 / 61.9	37.6 / 62.8
pl_en	32.1 / 59.5	30.5 / 58.6	31.8 / 59.4	31.5 / 59.3	8.5 / 32.3	25.7 / 54.6	26.2 / 55.4	26 / 55.4
ps_en	36.6 / 61.8	30.1 / 57.1	30.6 / 57.2	32.6 / 59.1	1.2 / 11.8	15.7 / 39.4	22.9 / 49	23.2 / 50.7

pt_en	51.7 / 72.7	50.7 / 72.7	52.3 / 73.6	51.7 / 73.2	5.5 / 24	43.4 / 67.8	44.2 / 68.4	44.3 / 68.3
quy_en	13.6 / 36.6	8.7 / 31.5	10.3 / 32.7	11.4 / 34.1	0.3 / 8.5	2.8 / 23.2	2.7 / 21.7	3.5 / 22.8
rn_en	27.4 / 51.2	19.9 / 43.2	23.5 / 46.6	23.6 / 46.9	0.2 / 7.1	10.1 / 30.6	12.3 / 35.4	11.6 / 35.3
ro_en	46.9 / 70.4	45.1 / 69.6	46.2 / 70.3	46.1 / 70.2	6.3 / 27.3	37.6 / 62.9	38.2 / 64.9	39 / 64.9
ru_en	38.5 / 63.7	37.5 / 63.1	38.8 / 64.1	38.3 / 63.8	3.2 / 15.8	29.3 / 55.7	31.1 / 58.9	30.5 / 59
rw_en	34.8 / 57.4	29.1 / 52.1	33.5 / 55.9	33.7 / 56.2	1.9 / 18.2	19.8 / 43.2	19.1 / 42.5	17.3 / 44.9
sa_en	26.1 / 52.9	20.7 / 47.4	21.5 / 48	16.7 / 41.8	1.3 / 12.3	11.8 / 37.9	13.6 / 42.3	11.4 / 42.5
#N/A sc_en	47.8 / 69.3	37.1 / 62.1	30.1 / 56.6	36.9 / 62.3	3 / 22.8	18.4 / 45.6	20.6 / 48	20.5 / 49.2
scn_en	41.3 / 64.5	33.5 / 59.8	27.4 / 54.8	33.3 / 59.6	3.8 / 24.5	24.3 / 50.5	26.7 / 53.6	25.1 / 53.9
sd_en	45 / 67.9	39 / 62.9	42.2 / 65.6	41.7 / 65	1.1 / 9.8	24.2 / 50.5	25.3 / 51.8	26.4 / 53
sg_en	14.4 / 36.7	8.2 / 31.1	9 / 31.7	10 / 32.8	0.2 / 9.7	3.5 / 19.2	2.7 / 19.3	2.7 / 22.1
shn_en	26.5 / 51.5	18.6 / 45.3	17.3 / 43.5	20.2 / 47	0.1 / 3.3	1.9 / 17.3	2 / 18.4	1.8 / 17.4
si_en	37.9 / 63.7	34.1 / 60.4	36 / 61.7	35.9 / 62.2	0.7 / 6.3	17.2 / 45.6	21.6 / 50.2	21.9 / 50.6
sk_en	41.1 / 66.1	41.4 / 66.9	42.2 / 67.5	42.3 / 67.7	7.3 / 29.5	34.3 / 61.4	35.7 / 62.5	35.5 / 62.4
sl_en	36.5 / 63	37.1 / 63.2	37.8 / 63.9	37.9 / 63.9	4.8 / 25.4	31.5 / 58.8	32.1 / 59.4	31.7 / 59.1
sm_en	35.8 / 58.3	28 / 52.3	31.6 / 55.9	33.4 / 56.4	0.7 / 13.1	14.2 / 36.3	18.1 / 42	15.4 / 41.8
sn_en	28.9 / 51.9	23.7 / 46.6	27.2 / 49.9	26.9 / 49.9	0.6 / 11.4	12.4 / 33.2	13.1 / 37.5	14.7 / 37.7
so_en	30.2 / 54.4	27.8 / 51.6	30.8 / 54.6	31.2 / 55	1 / 11.6	17.5 / 40.6	21.1 / 45.1	20.1 / 45.7
sq_en	44.9 / 68.4	41.6 / 66.9	43.7 / 68.2	43.5 / 68	11.3 / 34.5	35.8 / 62	36.5 / 62.4	37.1 / 63.4
sr_en	46.3 / 69.4	44.1 / 68	46.3 / 69.4	45.7 / 69.2	6.1 / 22.2	38.7 / 63.8	39.1 / 65.4	40.7 / 66
ss_en	29.2 / 52.3	21.8 / 43.8	25.3 / 47.5	25 / 47.7	0.4 / 11	6.9 / 25.1	8.2 / 29.2	8.9 / 32.3
st_en	41.2 / 62.9	27.9 / 51.2	34.2 / 57	34 / 56.8	0.6 / 12	10.1 / 30.2	15.5 / 38.9	13.2 / 40.1
su_en	40.2 / 63.6	34.3 / 59.1	32.6 / 58.1	35.8 / 60.7	1.8 / 16.4	26.4 / 51.8	26.6 / 52.3	29.2 / 54.9
sv_en	49.4 / 71.2	49.8 / 71.6	50.4 / 72.1	50.4 / 72.2	7.6 / 30.5	44.7 / 67.8	45.3 / 68.1	45.5 / 68.4
sw_en	46.2 / 67.4	43.3 / 64.9	46.5 / 67.1	46.1 / 66.9	1.8 / 15.3	30.8 / 54.3	34.3 / 57.9	36 / 59.5
szl_en	45.8 / 69	39.4 / 64.4	34 / 60.4	40.3 / 65.2	6.9 / 29	29.6 / 54.9	29.2 / 55.6	28.8 / 56.2
ta_en	36.8 / 62.8	30.9 / 57.4	32.3 / 58	33.1 / 59.3	1 / 5.2	22.9 / 51	22.7 / 50.8	24.1 / 52.4
taq_Tfng_en	8.8 / 30.1	4.9 / 27.3	6.7 / 28.8	6.5 / 29.6	0.1 / 5.4	0.2 / 12.9	0.4 / 14.7	0.3 / 13.7
taq_en	13.4 / 35.8	10.2 / 32.8	9.6 / 32.6	11.5 / 34.7	0.3 / 9.4	1.1 / 18.7	2.1 / 18.6	3.3 / 20.1
te_en	43.6 / 67.3	37.4 / 62.4	38.2 / 62.5	39.5 / 63.9	1.1 / 3.3	23.8 / 51.5	27.5 / 55.9	27.2 / 55.4
tg_en	36.5 / 61.7	13.7 / 38.2	24.8 / 50.5	22.5 / 47.3	2.3 / 16.1	23.4 / 49.9	26.5 / 54.7	27.3 / 54.3
th_en	33 / 60	31.4 / 57.4	33.3 / 59.2	33.3 / 59.6	2.1 / 6.7	24.1 / 53.1	24.1 / 54.7	24.8 / 54.5
ti_en	26.4 / 52.4	15.2 / 41.5	19.8 / 46	20.6 / 47.1	0.6 / 10.2	8 / 29.8	8.8 / 34.6	9.3 / 35
tk_en	34.9 / 61.2	31.3 / 57.7	33 / 58.9	33.5 / 58.8	2.4 / 18.5	19.1 / 45.6	20 / 47	20.2 / 49.2
tn_en	29.7 / 53.4	18.7 / 42.6	21.9 / 46.7	23.4 / 46.8	0.3 / 10.8	7.4 / 25.9	9.9 / 31.8	7.4 / 31.1
tpi_en	30.4 / 54.6	10.7 / 37.8	10.5 / 37.3	14.2 / 42.2	0.9 / 15.8	7.3 / 31.1	8.3 / 28.7	8.9 / 29.3
tr_en	41.5 / 66.1	39.1 / 63.8	41.4 / 65.5	40.9 / 65.3	3.4 / 20.4	25 / 51.8	28.1 / 55.5	29.1 / 56.7
ts_en	32.1 / 54.9	19.6 / 42.3	24.3 / 47.3	25.4 / 47.7	0.4 / 10.3	7.4 / 25.9	8.8 / 28.1	7.8 / 30.7
tt_en	31.8 / 58.3	10.4 / 30.4	10.3 / 29.4	17.5 / 38.5	1.2 / 9.9	16.5 / 42.1	21.3 / 48.5	21.3 / 48.5
tum_en	20.5 / 44.9	12.2 / 37	15.2 / 40	15.9 / 40.5	0.2 / 6.9	7.2 / 27.2	7 / 27.3	7.2 / 28.7
ug_en	27 / 54.6	8.6 / 29	14.1 / 36	10.9 / 30.6	1.6 / 14.4	13.4 / 39.4	17.3 / 44.9	17.3 / 44.9
uk_en	42.4 / 66.7	42.1 / 66.4	43 / 66.8	42.5 / 66.7	6.6 / 24.8	34.6 / 60.5	35.9 / 61.7	34.5 / 62
umb_en	11.4 / 33.3	4.4 / 24.2	4.7 / 24.2	5.5 / 25.9	0.2 / 9.6	1.1 / 17.2	2.3 / 18.3	1.6 / 17.1
ur_en	39.6 / 64.8	33.3 / 59.6	36.3 / 62.1	35.9 / 61.7	3.1 / 17.4	24.8 / 52.8	26.7 / 54.1	28.7 / 56.2
uz_en	35.2 / 61.6	33.9 / 59	37.2 / 62.1	37.8 / 63.2	1.8 / 15.6	23.5 / 51.2	24.6 / 52.3	24.9 / 52.9
vec_en	45.4 / 68	39.4 / 64.6	33.4 / 59.9	39.4 / 64.7	4 / 23.6	31 / 56.5	34.1 / 60	35.1 / 60.2
vi_en	40 / 64.2	37.4 / 62.5	39.2 / 63.9	39.3 / 63.8	4.1 / 19.7	28 / 53.5	31.7 / 57.3	32.1 / 57.9
war_en	51 / 71.1	32.1 / 58.6	32.5 / 58.8	35.9 / 60.9	1.1 / 15.5	25 / 50	27.4 / 55.3	31.8 / 56.4
wo_en	19.8 / 42.9	6.6 / 28.4	6.7 / 29.3	7.1 / 29.8	0.3 / 9	1.8 / 16.9	3 / 18.2	3.2 / 19.7
xh_en	39.7 / 61.1	32.1 / 54.3	36.5 / 58.2	35.5 / 57.7	1 / 15.2	17.2 / 39.3	21 / 43.9	21.1 / 44

yi_en	53.6 / 72.4	38.8 / 62.6	48.8 / 69	45.5 / 67.4	1.2 / 8.5	34.2 / 56	37.6 / 59.6	39.1 / 62.2
yo_en	23.9 / 47.9	11.9 / 34.9	17.4 / 40.8	17.8 / 40.7	0.4 / 12.5	5.1 / 23.5	4 / 24.1	6.3 / 26
yue_en	30.8 / 58.5	30.4 / 57.9	31.7 / 58.7	32 / 59.3	0.3 / 1.3	21.5 / 49.9	22.5 / 51.3	23.3 / 52.5
zh_Hant_en	28.8 / 56.2	28.1 / 55.7	29.6 / 56.8	30 / 57.5	0.6 / 2.7	19.6 / 47.3	21.2 / 49.6	21.6 / 50.5
zh_en	31.9 / 59.5	29 / 57.9	30.4 / 59	30.5 / 59.2	0.3 / 1.9	21.8 / 50.2	22.5 / 51.5	23.4 / 52.1
zu_en	41.3 / 62.7	33 / 55.1	37.7 / 59.1	37.1 / 58.9	0.8 / 13.4	17.5 / 40	20.3 / 44	20.2 / 45.5
xx2en	**35.5 / 59.6**	29.7 / 54.4	30.9 / 55.4	31.9 / 56.4	2.0 / 13.3	20.5 / 44.1	22.3 / 46.9	22.4 / 47.6
en2xx	**20.7 / 50.**1	17.3 / 44.1	17.8 / 44.7	18.6 / 45.7	0.4 / 5.7	8.1 / 26.7	8.7 / 29.0	8.7 / 28.8
Mean	**28.2 / 54.9**	23.5 / 49.2	24.4 / 50.0	25.3 / 51.1	1.2 / 9.6	14.3 / 35.5	15.6 / 38.0	15.6 / 38.2
xx2yy	**13.7 / 40.5**	8.8 / 31.2	8.4 / 30.9	10.1 / 34.0	0.3 / 4.1	4.0 / 16.1	4.4 / 17.3	4.2 / 17.1

Table A.10: Evaluation scores on Flores-200 direct pairs (depicted as `<bleu> / <chrf>`) for the MT models and language models described in Section 2.1.4.1 and Section 2.1.4.2 compared against NLLB-54B. All metrics are computed with the sacrebleu reference implementation.

	NLLB	**MT-3B**	**MT-7.2B**	**MT-10.7B**	**LM-8B**			
					0-shot	1-shot	5-shot	10-shot
ckb_cs	20.7 / 47.3	15.4 / 39.4	18.5 / 44.9	19.1 / 45.1	0 / 0.8	8.1 / 30.2	10.4 / 35.5	9.7 / 36.6
ckb_cy	26.9 / 53.6	19.3 / 41.4	23.6 / 47	23.9 / 47.8	0 / 0.7	0.3 / 8.8	0.7 / 15.1	0.3 / 12.4
ckb_en	37.5 / 61.6	28.3 / 53.1	34.4 / 58.9	34.2 / 59	1 / 9.7	17.9 / 42.5	22.4 / 48.4	22.7 / 49.4
ckb_et	16.4 / 48.6	13 / 40.8	15.3 / 45.6	16.3 / 47	0 / 1	1.9 / 12.9	6.7 / 36.6	7.3 / 38.2
ckb_eu	11.3 / 47.4	7.2 / 38.6	10 / 44.2	11.5 / 46	0 / 0.5	1.3 / 18.4	1.9 / 30.8	2.1 / 33.5
ckb_fon	2.1 / 19.9	1 / 14.7	0.4 / 10.8	1 / 14.4	0 / 0.2	0.3 / 8.5	0 / 1.8	0 / 3.2
ckb_fr	31.2 / 56.9	23.8 / 45.4	30.5 / 53.5	28.7 / 51.9	0.1 / 2.5	13.5 / 38.4	18.7 / 45.2	18.8 / 45.6
ckb_ln	15.5 / 47.5	1.3 / 16	0.5 / 14.4	1.2 / 17.4	0 / 0.3	0.1 / 6.2	0.1 / 10.3	0.2 / 11.4
ckb_mr	11 / 43.3	3.3 / 22.5	1.4 / 22.4	4.7 / 27.9	0 / 0.4	0.6 / 4.7	0.6 / 21.5	0.8 / 24.8
ckb_ny	10.4 / 44.7	3.5 / 29.7	2.7 / 31	6.8 / 35.9	0.1 / 0.4	0.3 / 13.1	0.1 / 9.9	0.1 / 8.5
ckb_or	8.6 / 41.8	7 / 36	7.3 / 38.5	8.5 / 40.1	0 / 0.1	0 / 0.2	0 / 0.3	0 / 0.3
ckb_so	8.8 / 42.2	5 / 32.4	4.1 / 34.4	6.9 / 37.2	0 / 0.3	0.3 / 6.6	0.1 / 5.8	0.1 / 7.4
ckb_ss	6.5 / 44.3	2 / 27.4	0.9 / 23.5	3.1 / 30.9	0 / 0.2	0.1 / 7.7	0.1 / 9.8	0 / 9.3
ckb_ti	4 / 24.2	1.1 / 12.8	0.3 / 10.5	1.7 / 16	0 / 0.1	0 / 0.4	0 / 0.1	0 / 0.1
ckb_yo	5.5 / 27	1.2 / 15.2	1.2 / 16.6	1.8 / 18.2	0 / 0.3	0 / 4.1	0 / 3.3	0 / 4
ckb_zh	27.6 / 25.2	3 / 18.9	2.2 / 22.1	4.1 / 22.1	0.4 / 0.9	1.5 / 3.4	15.7 / 15.7	14.6 / 16
cs_ckb	8.8 / 45.4	5.8 / 37.6	1.3 / 30.8	6 / 37.8	0 / 0.2	0.1 / 8	0.1 / 9.9	0.1 / 8.8
cs_cy	29.7 / 57.6	24.3 / 48	28.4 / 52.4	27.4 / 51.9	0.3 / 7.4	2.5 / 19.3	3.3 / 29.4	2.2 / 28.5
cs_en	41.2 / 66.4	41.2 / 66.6	42.3 / 67.4	42.3 / 67.3	7.9 / 28.9	34.7 / 61.4	35.9 / 62.6	36.2 / 62.9
cs_et	20.4 / 53.9	22.8 / 54.9	22.3 / 54.5	23.5 / 56	1.8 / 14.1	15.2 / 45.8	17.9 / 51.1	17.6 / 51.3
cs_eu	11.8 / 47.8	11.9 / 45.3	8.9 / 44.7	12.6 / 47.5	0.2 / 7.6	7.3 / 40.6	6.3 / 44.6	7 / 46
cs_fon	2.5 / 21.6	1.2 / 14.3	0.7 / 11.9	1 / 14	0 / 2.3	0.1 / 4.3	0 / 5	0 / 1.7
cs_fr	35.3 / 60.9	38.7 / 62.3	39.5 / 62.6	40.1 / 63.2	1.2 / 10.2	24.5 / 51.6	24.2 / 54	25.6 / 54.6
cs_ln	15.6 / 47.8	1.6 / 17.6	1.1 / 17.1	2.3 / 20.9	0.2 / 6.2	0.3 / 10.6	0.1 / 8.9	0.1 / 7.2
cs_mr	10.8 / 43.1	3.6 / 24.9	1.6 / 22.9	4.5 / 26.8	0.2 / 2.7	3.7 / 28.8	3.5 / 33.8	2.9 / 35.2
cs_ny	10.1 / 44.7	7.3 / 38.3	3.9 / 36.2	7.9 / 39.3	0.2 / 5.3	0.2 / 16	0.3 / 13.6	0.1 / 12.1

cs_or	9.1 / 42.1	8.7 / 40.2	9.3 / 41.3	9.9 / 42.5	0.1 / 0.2	0 / 4.9	0.2 / 0.5	0.1 / 0.5
cs_so	9.1 / 42.7	7.7 / 39	7.3 / 39.3	7.7 / 39.6	0.2 / 5	0.4 / 14.3	0.2 / 13.5	0.1 / 10.9
cs_ss	7.3 / 45.1	2.4 / 26.2	1.2 / 25.4	3.2 / 30.4	0.2 / 6.3	0.2 / 20.7	0 / 6.3	0 / 7.4
cs_ti	4.3 / 24.9	1.4 / 12.7	0.3 / 9.8	1.6 / 16.1	0 / 0.3	0 / 0.7	0 / 0.7	0 / 0.4
cs_yo	4.5 / 25.1	2 / 18.4	1.4 / 16.3	2 / 17.7	0.1 / 4.8	0.1 / 11.7	0 / 2.1	0 / 7.1
cs_zh	23.6 / 23.3	6.4 / 31.1	7.3 / 31.8	6.8 / 32.7	2 / 2.3	20.2 / 19.9	19.7 / 20	21.6 / 20.5
cy_ckb	10.5 / 48.3	8 / 43.1	5.2 / 41.6	8.4 / 42.7	0 / 0.2	0 / 5.6	0 / 5.2	0 / 2.1
cy_cs	27.2 / 54.1	28.9 / 55.6	29.8 / 56.6	29.7 / 56.4	0.3 / 4.7	16.1 / 41.8	16 / 43.4	12.8 / 43.4
cy_en	60 / 77.1	57.6 / 75.6	60.7 / 77.7	60.8 / 77.7	5.4 / 24.3	43.9 / 65.6	48.5 / 69.3	47.7 / 68.7
cy_et	22.3 / 55	24.2 / 56.4	25 / 57.5	25.8 / 57.9	0.1 / 4.8	5.6 / 31.3	7.6 / 40.8	7.5 / 42.2
cy_eu	14.4 / 51.6	17.1 / 54.9	19 / 56.7	18.5 / 56.3	0 / 4.8	0.6 / 16.8	2.8 / 36.2	3.3 / 37
cy_fon	2.2 / 18.6	1.3 / 14.9	0.6 / 11.7	1.3 / 14.9	0 / 0.9	0.5 / 7.1	0 / 1.7	0 / 5.5
cy_fr	40.9 / 64.2	44 / 65.6	45.8 / 66.7	45.3 / 66.7	0.7 / 9.3	27.7 / 53.8	28.7 / 54.9	27.4 / 55.7
cy_ln	17.4 / 49.1	2 / 18	1.1 / 17.5	2 / 20.1	0.1 / 3.8	0.3 / 11.4	0.1 / 8.1	0.1 / 8.8
cy_mr	13.9 / 47.5	10.2 / 41.3	6.4 / 39.1	11.1 / 42.4	0.1 / 1.5	2.8 / 26	2 / 29.7	1 / 26.5
cy_ny	12.4 / 46.5	6.1 / 36.2	5.6 / 40.4	9.3 / 41.9	0.1 / 3	0.8 / 13.6	0.2 / 11	0.1 / 9.1
cy_or	11.5 / 45.1	11.2 / 44.4	12.1 / 46	12.3 / 46.2	0 / 0.3	0 / 0.4	0.1 / 0.4	0.1 / 0.4
cy_so	10.8 / 44.8	8.6 / 41.3	8.2 / 42.4	9.6 / 43	0.1 / 3.4	1 / 13.3	0.4 / 14.5	0.2 / 12.9
cy_ss	9 / 46.7	3.2 / 31.3	1.4 / 28.3	4.1 / 34	0.1 / 3	0.2 / 12.3	0.4 / 11.3	0.2 / 11.4
cy_ti	4.8 / 25.5	1.5 / 14.2	0.4 / 12.3	2.1 / 17.9	0 / 0.3	0 / 0.8	0 / 0.5	0 / 0.2
cy_yo	5.4 / 26.3	2.1 / 17.7	1.7 / 18.3	2.2 / 19.4	0.1 / 2.5	0 / 2.9	0 / 4.1	0 / 6.4
cy_zh	30.4 / 28.1	5.4 / 31.5	5.7 / 33.1	5.3 / 32.9	0.8 / 1.4	8.4 / 9	19.5 / 20.1	15.3 / 19.3
en_ckb	13.2 / 52.5	9.1 / 44.9	4.5 / 42.8	9.4 / 44.4	0.1 / 0.3	0.3 / 10.2	0.1 / 13.8	0.1 / 13.4
en_cs	33.6 / 59.8	34.8 / 61.2	34.8 / 61.1	35.6 / 61.5	1.1 / 11.1	25.3 / 53.1	26.2 / 54.5	25.6 / 54.4
en_cy	51.6 / 72.5	41.6 / 62.4	45.6 / 66.3	44.7 / 65.6	0.5 / 9	6.4 / 29.4	8.2 / 38	4.8 / 35.4
en_et	27.1 / 59.9	28.9 / 61.3	28.1 / 61.2	29.8 / 61.8	0.9 / 11.9	22.2 / 54.7	22.8 / 55.6	22.2 / 55.3
en_eu	17.2 / 54.9	18.3 / 55.5	16.8 / 55.4	19.3 / 57.2	0.2 / 11.1	12.3 / 49.3	12.9 / 51	12.4 / 51.3
en_fon	3.1 / 23.5	1.6 / 15.6	1.3 / 14.6	1.6 / 15.9	0 / 0.3	0.1 / 4.2	0 / 1.9	0 / 1.6
en_fr	51.4 / 71.4	52.4 / 72	53.3 / 72.4	53.2 / 72.5	1.7 / 15.3	37.1 / 61.9	38.9 / 63.5	38.8 / 63.1
en_ln	19.1 / 50.8	2.2 / 18.3	1.5 / 18.2	2.7 / 20.8	0.4 / 9.3	0.7 / 11.3	0.1 / 9.2	0.2 / 10.7
en_mr	17.6 / 52.1	10.1 / 39	5.8 / 36.4	11.8 / 42	0.8 / 12	10.5 / 42.8	9 / 43.2	8.1 / 43.7
en_ny	13.6 / 48.7	8.2 / 38.7	6.6 / 43.3	10.6 / 43.5	0.4 / 8	0.7 / 16.4	0.4 / 16	0.4 / 17.6
en_or	15.1 / 50.4	12.4 / 46.8	13.2 / 47.4	13.1 / 47.8	0.1 / 0.2	0.4 / 0.9	0.3 / 0.8	0.3 / 0.8
en_so	12.6 / 47.3	9.4 / 41.8	9.9 / 43.2	10.2 / 43	0.2 / 6.9	0.6 / 15.2	0.2 / 13	0.2 / 13.3
en_ss	10.7 / 49.2	4.1 / 33.5	1.9 / 31.9	5.3 / 36	0.5 / 7.2	0.6 / 15	0.1 / 11.6	0.1 / 9.5
en_ti	6 / 27.9	1.9 / 15.7	0.6 / 14.6	2.5 / 18.9	0.1 / 0.3	0 / 0.9	0 / 0.8	0 / 0.6
en_yo	6 / 27.4	2.6 / 19.7	1.9 / 19.6	2.9 / 20.7	0.3 / 6.3	0.1 / 10.6	0.1 / 5.3	0 / 3.3
en_zh	31.2 / 29.3	5.9 / 35.7	6 / 36.5	5.8 / 36.5	2.9 / 3.3	29.1 / 26.1	29.9 / 26.6	29.9 / 26.3
et_ckb	8.7 / 44.9	4.9 / 34.9	2.8 / 34.8	5 / 34.7	0 / 0.4	0.1 / 10.4	0 / 8.7	0 / 5.9
et_cs	24.1 / 51.6	25.3 / 52.3	24.9 / 51.8	25.3 / 52.6	1.8 / 14.1	15.8 / 42	18.2 / 46	17.4 / 46
et_cy	27.9 / 56.1	23.2 / 47.2	25.4 / 49.3	25.9 / 49.9	0.4 / 10.2	2.5 / 22.3	1.5 / 25	1.2 / 23.3
et_en	38.8 / 64.6	39.4 / 64.8	40.1 / 65.5	39.6 / 65.5	7.7 / 27.9	31.8 / 59.1	32.5 / 59.8	32.6 / 59.7
et_eu	11.4 / 47	13.4 / 48.5	13.5 / 49.1	14 / 50	0.3 / 12.2	5.9 / 37.8	5.7 / 41.2	3.9 / 41.1
et_fon	2.5 / 22.4	1 / 13.2	0.5 / 11.3	0.9 / 12.7	0 / 2.6	0.1 / 4.8	0 / 3.5	0 / 3.5
et_fr	33.4 / 59.2	35.1 / 59.3	35.7 / 59.3	36 / 60.1	1.4 / 14.2	23 / 50.2	24.8 / 52.1	25.3 / 52.6
et_ln	14.8 / 47.6	1.7 / 17.3	0.8 / 16.1	2 / 20.8	0.3 / 10.7	0.2 / 9.5	0.1 / 9.3	0.1 / 10.5
et_mr	10.8 / 42.8	5.4 / 29.3	2.4 / 25.8	5 / 28.5	0.3 / 1.6	0.7 / 2.7	3.9 / 33.5	2.8 / 35.3
et_ny	9.7 / 43.8	6.1 / 34.2	3 / 33.2	7.8 / 37.4	0.4 / 9.9	0.1 / 7	0.1 / 11.4	0.1 / 7.9
et_or	9 / 41.8	8.2 / 39.5	7.7 / 39.9	9 / 41	0 / 0.3	0.1 / 1.8	0.3 / 0.5	0.3 / 0.5
et_so	8.6 / 42.4	7.6 / 37.8	6.6 / 38.4	8.1 / 39.5	0.3 / 9	0.5 / 15	0.4 / 14.5	0.1 / 12.2

et_ss	6.6 / 44.8	2.5 / 26.3	0.9 / 23.6	3.1 / 29.3	0.1 / 10.7	0.1 / 21.8	0 / 7.4	0 / 7.8
et_ti	4 / 24.1	1.2 / 11.6	0.3 / 8.9	1.3 / 13.8	0 / 0.3	0 / 0.8	0 / 0.7	0 / 1.2
et_yo	4.1 / 24.6	2 / 17.4	1 / 15	1.6 / 16.2	0.2 / 7.3	0 / 4.1	0 / 5.3	0 / 3.2
et_zh	22.9 / 22.6	6.3 / 30.9	6.1 / 31.7	5.7 / 31.6	2.1 / 2.8	21.2 / 20.1	18.9 / 21.3	22.7 / 22.8
eu_ckb	7.3 / 43.7	4.6 / 34.2	1 / 27.8	4.3 / 32.4	0 / 0.2	0.1 / 8.3	0 / 8.6	0 / 5.3
eu_cs	19.2 / 46.9	16.7 / 42.2	14.4 / 41.9	18.2 / 44.4	0.9 / 12.2	13 / 40.6	13.8 / 41.6	13.2 / 41.2
eu_cy	24 / 53.1	22 / 46.1	25.6 / 50.5	24.5 / 49.7	0.2 / 10.6	1.2 / 18.7	1.5 / 22.7	0.9 / 21.6
eu_en	34.4 / 61.2	32.6 / 58.6	34.1 / 59.8	34.4 / 60.4	4.6 / 23.3	25.5 / 53.8	24.7 / 52.9	26.2 / 54.4
eu_et	16.8 / 49.6	16.4 / 47.7	17.9 / 49.4	18.7 / 50.2	0.1 / 12.2	7.1 / 35.2	9.9 / 43.9	9.6 / 43.6
eu_fon	2.5 / 20.7	1 / 13.4	0.4 / 10.1	1 / 13	0.2 / 7.3	0.1 / 9.1	0 / 2.8	0 / 1.6
eu_fr	29.4 / 56.2	28.8 / 53	30.3 / 54	30.3 / 54.4	1.2 / 14	20.3 / 48.3	21.3 / 49.4	21.1 / 49.5
eu_ln	14.5 / 47	1.6 / 17	0.6 / 15.8	1.9 / 19	0.1 / 12.5	0.4 / 16.5	0.1 / 8.9	0.1 / 8.5
eu_mr	9.6 / 42.6	2.8 / 22.6	1.4 / 21	3.7 / 24.6	0.1 / 0.6	3.3 / 31	2.7 / 31.1	2.4 / 32.1
eu_ny	9.3 / 44.2	4.3 / 30.7	2.8 / 31.9	6.4 / 35.1	0.4 / 11.2	0.2 / 12.8	0.1 / 11.5	0.1 / 11.3
eu_or	7.9 / 40.3	8.1 / 39.6	7.4 / 40	8.6 / 41	0 / 0.3	0 / 3.3	0.2 / 0.5	0.1 / 0.5
eu_so	8.3 / 41.7	6.4 / 35.4	4.8 / 35.8	7.1 / 37.8	0.2 / 11.4	0.4 / 12.8	0.1 / 12	0.1 / 10.9
eu_ss	5.9 / 43.7	1.9 / 25	0.9 / 22.6	2.8 / 29.4	0.1 / 11.3	0.1 / 8.1	0 / 6	0 / 8.7
eu_ti	3.8 / 24.1	0.8 / 11	0.3 / 9.3	1.5 / 15	0 / 0.3	0 / 1.9	0 / 0.6	0 / 0.7
eu_yo	5 / 26.3	1.7 / 16.8	1 / 15.8	1.7 / 16.8	0.1 / 7.6	0.1 / 7.5	0 / 2.8	0 / 4.2
eu_zh	23.2 / 22.4	5.6 / 22.6	2.3 / 20.3	5.9 / 24.9	1.8 / 2.6	17.5 / 17	20.2 / 21	14.3 / 19.4
fon_ckb	2.8 / 28.6	0.9 / 20.6	0.3 / 16.8	1.6 / 24	0 / 0.2	0 / 0.2	0 / 2	0 / 1.5
fon_cs	7.6 / 27.8	2.8 / 20.2	2.1 / 19.3	4.3 / 22.3	0 / 1.3	0.5 / 13	0.8 / 13.9	0.3 / 13.5
fon_cy	8.8 / 31.5	4 / 21.8	2.9 / 21.2	6 / 25.6	0 / 2.2	0 / 7.3	0.1 / 10.4	0.1 / 7.1
fon_en	12.9 / 35.4	3.9 / 25	2.2 / 23.4	6.4 / 27.8	0.1 / 7.8	0.8 / 12.5	1.4 / 17.7	1.9 / 18.3
fon_et	6.7 / 30.6	2.2 / 22.3	1.2 / 20.4	3.4 / 24.9	0 / 1.3	0.7 / 15.5	0.4 / 15.8	0.6 / 16.1
fon_eu	4.6 / 31.3	1.4 / 22.5	0.9 / 19.8	2.9 / 27.2	0 / 2	0.5 / 17.6	0.1 / 12.4	0.2 / 14.3
fon_fr	10.6 / 33.2	5.3 / 24.4	4.4 / 24.3	7.6 / 26.8	0.1 / 3.7	0.9 / 11.2	1.3 / 17.1	1.3 / 16.4
fon_ln	8.2 / 36.8	0.6 / 11	0.2 / 7.6	0.5 / 11.8	0.1 / 3	0.1 / 6.6	0 / 5	0 / 1.5
fon_mr	3.8 / 27	0.8 / 16.5	0.2 / 13.4	1.5 / 18.7	0 / 0.4	0.3 / 8.3	0.1 / 8.7	0 / 7
fon_ny	5.6 / 33.3	1.3 / 20.6	0.7 / 18.7	3 / 26.5	0 / 1.9	0.1 / 12.8	0 / 4.4	0 / 2.7
fon_or	2.7 / 25.1	1.1 / 18.5	0.7 / 17	2 / 22	0 / 0.2	0 / 0.2	0 / 0.4	0 / 0.3
fon_so	4.3 / 31.2	1.6 / 23	0.8 / 19.6	2.4 / 26.3	0 / 2	0.1 / 5.6	0 / 5.5	0 / 4.1
fon_ss	3 / 32.5	0.8 / 16.7	0.3 / 14	1.8 / 23.2	0 / 2.4	0 / 4.5	0 / 3.9	0 / 1.8
fon_ti	1.7 / 15.5	0.3 / 6.5	0.1 / 5.5	0.7 / 10.2	0 / 0.2	0 / 0.2	0 / 0.1	0 / 0.1
fon_yo	3.4 / 21.3	0.7 / 12.3	0.4 / 10.8	1.1 / 14.9	0 / 2.3	0 / 3.2	0 / 3.7	0 / 1
fon_zh	11.8 / 12.9	1.9 / 8.7	0.5 / 6.2	2.2 / 10.2	0.1 / 0.6	1.6 / 4.3	1.8 / 5.1	0.7 / 3.5
fr_ckb	9.2 / 46.9	6 / 39.7	3.3 / 38.4	6.1 / 39.1	0 / 0.3	0.1 / 7.8	0.1 / 13.8	0.1 / 13.9
fr_cs	26 / 53.6	27.3 / 54.9	27.8 / 55.3	27.8 / 55.2	1.8 / 14.2	19.3 / 46.8	20.9 / 49.1	20.4 / 49.4
fr_cy	33.7 / 60.1	28 / 51.6	31.7 / 56	30.7 / 55	0.3 / 9.5	4.1 / 27.8	4.2 / 32.8	3.7 / 34.1
fr_en	47.9 / 70.1	46.1 / 69.6	47.3 / 70.3	47.1 / 70.1	2.7 / 21.4	39.2 / 64.5	39.9 / 64.9	40.1 / 65.1
fr_et	20.9 / 54.3	22.2 / 55.2	18.5 / 54.5	23.6 / 56.3	1.7 / 15.8	11.1 / 44.6	16.9 / 50.6	15.8 / 51
fr_eu	14 / 51.2	13.1 / 50.8	11.6 / 51.3	16.1 / 53.5	0.3 / 9.9	10.2 / 47.7	9.5 / 48.8	9.6 / 49.4
fr_fon	2.5 / 22.6	1.1 / 14.7	0.7 / 12.3	1.2 / 14.5	0.2 / 2.4	0 / 1.7	0 / 3.8	0 / 1.8
fr_ln	16.8 / 49.3	1.7 / 17.9	1.2 / 17.3	2.1 / 20.2	0.2 / 7.3	0.1 / 8.7	0 / 6.3	0.2 / 10.6
fr_mr	12 / 45.5	7.1 / 34.8	2.2 / 27	6.7 / 33.2	0.5 / 8.8	5.5 / 35.9	5.3 / 38.3	4.5 / 38.4
fr_ny	10.8 / 45.5	5.4 / 34.9	4.1 / 38.3	7.8 / 40	0.2 / 6.6	0.5 / 16.1	0.2 / 13.8	0.2 / 13.8
fr_or	10.2 / 43.8	9.3 / 41.7	10.1 / 43.5	10.1 / 43.6	0 / 0.2	0.1 / 1.1	0.1 / 0.6	0.2 / 0.6
fr_so	9.7 / 43.8	7.5 / 39.5	6 / 39.9	7.7 / 40.1	0.1 / 5	0.5 / 15.9	0.3 / 15	0.1 / 11.1
fr_ss	7.9 / 45.5	2.4 / 28.5	1 / 26.5	3.4 / 31.9	0.1 / 6.9	0.1 / 10.3	0.1 / 10.2	0.1 / 8.9
fr_ti	4.6 / 25.3	1 / 12.9	0.5 / 11.7	1.8 / 16.9	0 / 0.3	0 / 1.1	0 / 0.8	0 / 0.5

fr_yo	4.9 / 26.2	2.1 / 18.7	1.3 / 17.6	1.9 / 18.8	0.1 / 4.6	0.1 / 6.6	0 / 3.8	0 / 10.2
fr_zh	25.5 / 24.4	7.4 / 33	7.2 / 34.1	6.8 / 33.6	4 / 5.5	16.2 / 18.1	22.8 / 23.6	19.3 / 21.8
ln_ckb	6.6 / 40	1.1 / 22.1	0.3 / 18	1.4 / 22.8	0 / 0.2	0 / 3	0 / 3.6	0 / 3.8
ln_cs	15.6 / 40.7	3.6 / 22.1	1.4 / 18.1	4.7 / 23.6	0.2 / 4.7	2.6 / 19.2	2.5 / 21.2	1.9 / 20.4
ln_cy	20.2 / 45.8	4.9 / 23.2	2.6 / 21	6.4 / 26.6	0.1 / 5.2	0.1 / 5.8	0.1 / 7.8	0.1 / 6.7
ln_en	28.9 / 52.1	5.5 / 27.2	1.7 / 23.4	6.8 / 29.5	0.3 / 11.1	6.5 / 27.2	6.6 / 25.5	6.7 / 29.9
ln_et	12.8 / 42.6	3.2 / 24.8	1.3 / 20	3.7 / 26.3	0.2 / 5.2	0.3 / 12.9	2.3 / 23.7	1.9 / 22.9
ln_eu	9.3 / 42.1	2.1 / 25.3	1.3 / 22	3.2 / 28.3	0.1 / 5.9	0.7 / 17.7	1.4 / 25.3	0.6 / 22.3
ln_fon	2.6 / 22.6	1.1 / 13.1	0.3 / 8.4	0.8 / 13.3	0 / 0.5	0.1 / 5.4	0 / 3	0 / 1.8
ln_fr	23.8 / 48.9	6.5 / 25.5	2.8 / 24	7.6 / 28	0.4 / 9.9	4 / 23.1	6.5 / 30.9	4.1 / 28
ln_mr	8.1 / 37.1	1.2 / 16.8	0.3 / 12.8	1.5 / 18.7	0.1 / 0.7	0.6 / 16.4	0.1 / 10.1	0.2 / 13.6
ln_ny	9.4 / 42.1	1.6 / 20.5	0.6 / 17.3	2.6 / 25.9	0.2 / 6.4	0.1 / 8.2	0.1 / 9.6	0.1 / 8.9
ln_or	6.8 / 36.6	1.4 / 18.9	0.5 / 15.9	1.8 / 21.7	0 / 0.3	0 / 0.3	0 / 0.4	0 / 0.4
ln_so	8.6 / 40	2.1 / 24	0.7 / 19.4	2.6 / 26.6	0.1 / 5.9	0.4 / 12.9	0.1 / 9.8	0.1 / 8.9
ln_ss	7.1 / 43	1.2 / 18.5	0.4 / 13.6	2 / 22.5	0.1 / 5.2	0.1 / 9.5	0.1 / 8.3	0 / 7.2
ln_ti	3.4 / 20.6	0.3 / 4.5	0.1 / 5.7	0.6 / 10	0 / 0.3	0 / 0.2	0 / 0.2	0 / 0.2
ln_yo	4.3 / 24.5	0.8 / 12.9	0.4 / 11	1.1 / 14.9	0.1 / 4	0 / 1.4	0.1 / 7	0.1 / 7.4
ln_zh	20 / 19.9	1.7 / 8.9	0.3 / 4.3	2.2 / 10.7	0.3 / 1.2	1.5 / 5.1	4.2 / 8.6	2.9 / 7.9
mr_ckb	8.1 / 45.2	4.4 / 35.5	1.1 / 28.5	4.9 / 34.8	0.1 / 0.3	0 / 8.8	0 / 8.7	0 / 8
mr_cs	21.8 / 49.3	9.5 / 31.4	3.4 / 29.2	14.4 / 37.7	0.2 / 0.8	11.7 / 36.7	12.5 / 38.7	10.8 / 38.9
mr_cy	28.1 / 56.1	22.3 / 47	25.6 / 51.2	26 / 51.1	0.1 / 1	2.1 / 21.1	1.3 / 23.6	0.9 / 21.2
mr_en	40.3 / 65.8	35 / 61.2	35.4 / 61.3	36.9 / 62.4	1.5 / 7.6	24.8 / 53.7	27.8 / 56.4	27.9 / 56.3
mr_et	18 / 51.3	7.5 / 34	7.6 / 41	15.6 / 45.7	0.2 / 1.7	10 / 41.6	8.7 / 43	9.7 / 44.2
mr_eu	11.3 / 48.3	6.3 / 35.8	3.7 / 35.6	9.6 / 42	0 / 0.8	4.4 / 36.5	3.4 / 39	2.3 / 37.6
mr_fon	2.3 / 22.2	0.9 / 13.2	0.3 / 8.9	0.8 / 12.6	0 / 0.2	0 / 3.6	0 / 5.1	0 / 1.2
mr_fr	32.2 / 58.7	18.5 / 39.8	10.7 / 39.2	24.7 / 47.7	0.5 / 1.4	17.2 / 44.8	14.3 / 45	18.2 / 47.3
mr_ln	15.3 / 47.8	1 / 14	0.3 / 13.5	1.7 / 19	0 / 0.3	0 / 6.5	0.1 / 8.2	0.1 / 6.9
mr_ny	9.9 / 44.7	2.9 / 28.3	2.9 / 34.6	7.1 / 38	0.1 / 1.3	0.2 / 11.6	0.1 / 10.1	0 / 7.8
mr_or	11 / 45.6	9.2 / 42.3	9.7 / 43.2	9.5 / 43.4	0.1 / 0.3	0.2 / 0.7	0.2 / 1.2	0 / 1
mr_so	8.8 / 42.8	6.2 / 36.7	3.5 / 36.3	7 / 39.1	0.1 / 1	0.2 / 10	0.1 / 7.9	0 / 7.3
mr_ss	7.1 / 44.9	1.8 / 24.7	0.6 / 20.8	3.2 / 30.8	0.1 / 0.6	0.1 / 11.7	0 / 6.6	0 / 5.8
mr_ti	4.5 / 25.2	0.7 / 10.6	0.4 / 11.1	1.7 / 16.2	0 / 0.2	0 / 1.4	0 / 0.8	0 / 0.6
mr_yo	5.1 / 26.5	1.8 / 17.5	0.9 / 16.4	1.7 / 18.1	0.1 / 0.4	0.1 / 7.1	0 / 3.1	0 / 3.9
mr_zh	25.4 / 24	4.8 / 19.3	1 / 13.6	5.6 / 23.3	0.1 / 0.4	17.7 / 17	19.9 / 19.4	14.8 / 18.9
ny_ckb	6.6 / 39	4.5 / 33.5	1.4 / 29.5	5.2 / 35.5	0 / 0.2	0 / 0.5	0 / 1.6	0 / 2.9
ny_cs	15.2 / 40.3	14.2 / 37.4	15.2 / 39.3	15.3 / 39.2	0.2 / 4.2	0.6 / 15.9	3.5 / 23.3	3.1 / 23.2
ny_cy	20.7 / 46	17.4 / 39.1	20.3 / 42.2	20 / 42.7	0.1 / 4.2	0.2 / 8.6	0.1 / 7.8	0.1 / 6.3
ny_en	29.2 / 52.9	24.2 / 47.2	24 / 49.7	27.2 / 50.5	0.6 / 12.2	8.4 / 27.7	12.4 / 35.5	13.9 / 37.1
ny_et	12.7 / 42.6	11.8 / 38.6	10.9 / 40.2	13.5 / 41.9	0.1 / 4.7	1.5 / 20.7	1.6 / 24.5	2 / 26.7
ny_eu	8.9 / 41.1	8.6 / 39	5.9 / 36.9	9.8 / 42.7	0 / 4.5	0.6 / 20.4	0.8 / 25.4	0.6 / 23
ny_fon	2.5 / 21.9	1.3 / 14.6	0.6 / 12.1	1.4 / 15.6	0 / 0.3	0 / 4.4	0 / 1.4	0 / 3.7
ny_fr	22.5 / 47.4	19.8 / 42.1	22.4 / 45.2	23.8 / 46.3	0.2 / 5.5	8.6 / 31.6	5.6 / 32.2	4.5 / 32.2
ny_ln	13.6 / 44.7	1 / 13.7	0.5 / 13	1.8 / 19.2	0.1 / 5.2	0.5 / 13.4	0.1 / 10.8	0.1 / 9.4
ny_mr	7.4 / 36.1	4.3 / 27.3	1.4 / 21.6	5.3 / 29.9	0 / 0.5	0.3 / 13.6	0.5 / 19.6	0.3 / 15.8
ny_or	6.2 / 35.9	5.3 / 33.9	6.6 / 35.4	6.7 / 36.4	0 / 0.3	0.1 / 0.2	0 / 0.3	0.1 / 0.5
ny_so	8.2 / 39.6	6.2 / 35.3	4.5 / 35	6.9 / 37.1	0.1 / 5.5	0.4 / 12.8	0.2 / 12.3	0.1 / 9.5
ny_ss	7.3 / 42.5	2.5 / 27.2	0.9 / 22.1	4.2 / 32.2	0.1 / 4	0.1 / 10.5	0.2 / 13.5	0.1 / 9.5
ny_ti	3.3 / 19.4	0.8 / 10.8	0.3 / 9.3	1.7 / 15.4	0 / 0.3	0 / 0.3	0 / 0.2	0 / 0.1
ny_yo	4 / 23.3	2 / 17.7	1.4 / 16.6	2.3 / 19	0.1 / 3.5	0.1 / 4.1	0 / 3.8	0 / 4.4
ny_zh	18.8 / 18.9	5 / 20.7	1.3 / 14.9	4.7 / 21.9	0.3 / 1.1	3.9 / 6.7	6.2 / 11.3	6 / 11.5

or_ckb	8.1 / 45.1	6 / 38.9	5.7 / 40.4	7.5 / 42	0 / 0.2	0 / 0.2	0 / 0.3	0 / 0.3
or_cs	21.3 / 49.1	20.3 / 47.5	21.6 / 47.9	21.7 / 48.9	0 / 0.7	8.1 / 32.7	9.5 / 36.9	10.6 / 36.7
or_cy	28.6 / 56.4	22.4 / 48	27.4 / 52.1	26.3 / 51.6	0.1 / 0.9	0.5 / 8.6	0.4 / 5.6	0.8 / 10.5
or_en	41.6 / 66.2	35 / 61	36.3 / 61.7	36.1 / 61.7	1 / 8.6	18.9 / 47.3	20.6 / 49.1	22.5 / 50.7
or_et	17.1 / 50.7	17.6 / 49.6	18.2 / 50.4	18.7 / 51.1	0.1 / 0.9	6.8 / 35	9.1 / 41.1	7.9 / 38.3
or_eu	11.9 / 48.8	13.3 / 49.9	14.1 / 51.1	13.9 / 51.4	0 / 0.5	2.1 / 24.3	1.9 / 28.8	2.2 / 36.6
or_fon	2.3 / 21.4	1 / 13.5	0.5 / 10.7	1 / 13.9	0 / 0.2	0 / 0.5	0 / 2.3	0 / 1.5
or_fr	32.3 / 58.7	29.1 / 54.1	31.3 / 55.7	32.6 / 57	0.1 / 1	12.8 / 39.4	13.8 / 42.6	16.5 / 44.6
or_ln	15.9 / 47.8	1.4 / 14.8	0.7 / 14.7	1.6 / 16.3	0 / 0.4	0.3 / 5.1	0.1 / 6.4	0.1 / 7.3
or_mr	12.8 / 47	9.1 / 42.1	9.9 / 42.8	10.3 / 43.2	0.2 / 4.2	4.5 / 27.6	2.4 / 34.6	6.4 / 40.1
or_ny	10.8 / 45.7	5.3 / 34.9	5.4 / 38.7	8 / 39.6	0.1 / 0.7	0.1 / 10	0.1 / 5.8	0.1 / 7.7
or_so	8.8 / 42.5	6.3 / 36.8	7.2 / 38.5	7.5 / 40	0.1 / 0.5	0.3 / 2.5	0 / 6.7	0.1 / 6.9
or_ss	7.1 / 44.6	2.3 / 26.6	1.5 / 26.3	3.4 / 32.2	0 / 0.4	0 / 3.7	0 / 4.7	0.1 / 7.6
or_ti	4.4 / 25.3	0.8 / 10.5	0.5 / 12.6	1.8 / 16.2	0 / 0.2	0 / 0.3	0 / 0.3	0 / 0.2
or_yo	5.6 / 27.3	1.6 / 17.2	1.4 / 17.4	1.9 / 18.7	0 / 0.4	0 / 3.3	0 / 2.9	0 / 2.8
or_zh	26 / 24.3	5.2 / 26.7	4.2 / 26.8	6.6 / 29.5	0.4 / 0.6	8.6 / 8.4	18.2 / 16.7	18 / 17.9
so_ckb	6.3 / 40.7	3.5 / 30	2 / 31.4	4.5 / 33.7	0 / 0.5	0.1 / 5.9	0 / 1.2	0 / 2.6
so_cs	15.6 / 41	14.6 / 40.1	16.4 / 41.9	16.4 / 42.2	0.1 / 3.4	5.6 / 24.5	6.5 / 29.2	5.8 / 31.1
so_cy	21.2 / 47	20.1 / 43	23.1 / 46	22.6 / 46.1	0.2 / 4.4	0.7 / 11.5	0.2 / 10.6	0.2 / 9.6
so_en	30.2 / 54.4	27.8 / 51.7	30.7 / 54.5	31.2 / 55	1 / 11.4	17.5 / 40.5	21.1 / 45.2	19.5 / 45.7
so_et	13.1 / 43.3	14 / 43	13.9 / 43.4	15 / 44.5	0.3 / 7.2	2.5 / 22	3.7 / 30.8	2.9 / 30.4
so_eu	9.1 / 42.4	9.8 / 42.2	9.4 / 42.1	11.4 / 45.4	0.1 / 4.8	0.7 / 18.3	0.5 / 21.5	0.1 / 13.3
so_fon	2.1 / 20	1 / 13.6	0.5 / 10.6	0.9 / 13.8	0 / 0.6	0.1 / 7.4	0 / 5.7	0 / 1.8
so_fr	24.5 / 49.8	24 / 47.8	26.4 / 49.7	26.1 / 49.7	0.4 / 7.1	13.8 / 38.9	11.3 / 39.6	11.3 / 40.2
so_ln	15.3 / 45.9	1.4 / 15.7	0.6 / 13.5	1.8 / 18.9	0.1 / 4.8	0.2 / 11.8	0.1 / 9.4	0.1 / 8.7
so_mr	8.7 / 38.8	4.8 / 28.9	1.3 / 21.5	5.1 / 28.5	0.2 / 6.1	1.8 / 23.7	1.1 / 24	0.7 / 22.6
so_ny	10.5 / 43.4	4.7 / 31.9	2.5 / 31.9	7 / 37.2	0.2 / 5	0.5 / 17.3	0.2 / 12.8	0.1 / 10.6
so_or	7 / 37.3	6.2 / 35	7.1 / 36.9	7.5 / 37.8	0.1 / 0.3	0.1 / 0.2	0.1 / 0.4	0.1 / 0.4
so_ss	6.6 / 42.4	2 / 25.5	0.8 / 21.4	3.5 / 31.2	0 / 3.6	0 / 12.9	0.2 / 12.3	0 / 6.5
so_ti	3.6 / 22.5	1 / 11.3	0.3 / 9.6	1.4 / 15.2	0 / 0.3	0 / 0.4	0 / 0.3	0 / 0.2
so_yo	6.1 / 27	1.5 / 16.2	1 / 15.7	1.8 / 17.9	0.1 / 3.2	0.1 / 11.9	0.1 / 9.2	0 / 6
so_zh	20.3 / 20.5	3.6 / 21.4	4.3 / 22.9	3.5 / 23	0.6 / 1.5	8.5 / 12.4	10.1 / 14.6	10.2 / 14.4
ss_ckb	6.2 / 39.2	4 / 32.4	1.8 / 30.2	4.5 / 33.3	0 / 0.2	0 / 3	0 / 1.7	0 / 1.9
ss_cs	14.9 / 39.1	11.1 / 33.3	13 / 36	12.8 / 36.2	0.1 / 3.6	0.7 / 13.5	1.7 / 20.4	1.3 / 20
ss_cy	20.1 / 45.4	15.8 / 36.8	16.7 / 39.1	18.8 / 40.7	0.1 / 4.5	0.4 / 11	0.1 / 8.6	0.1 / 6.9
ss_en	29.2 / 52.3	21.6 / 43.8	22.9 / 47	25 / 47.7	0.4 / 11	7 / 25.1	8.4 / 29.3	9.2 / 32.5
ss_et	12.1 / 41.6	9.6 / 35.5	8.8 / 37.2	11.3 / 39.1	0.1 / 4	0.2 / 14.9	1.2 / 20.7	1 / 20.3
ss_eu	8.2 / 39.7	7.7 / 36.8	7.1 / 37.4	8.9 / 40.5	0 / 4.9	0.1 / 10.7	0.4 / 20.9	0.3 / 18.7
ss_fon	2.7 / 22	1.4 / 14.8	0.6 / 11.5	1.4 / 14.9	0 / 0.3	0 / 3.4	0 / 3.9	0 / 1.5
ss_fr	20.7 / 45.4	18.3 / 40.1	21.3 / 43.1	20.9 / 43.3	0.2 / 5.9	4.4 / 25.5	4.1 / 27.3	2.5 / 26.8
ss_ln	12.1 / 42.8	1 / 14.4	0.5 / 11.6	1.5 / 18	0.1 / 6.4	0.1 / 10.3	0.1 / 9.9	0.1 / 8.7
ss_mr	6.7 / 34.7	3.4 / 24.7	1.3 / 20.4	4.4 / 27.4	0.1 / 0.9	0.4 / 13.5	0.3 / 16.6	0.2 / 13.3
ss_ny	7.6 / 38.4	5.1 / 31.9	3.2 / 31.8	7.7 / 36.9	0.2 / 6.8	0.1 / 9.2	0.1 / 9.6	0.1 / 9.6
ss_or	6.1 / 34.7	4.9 / 31.1	4.9 / 32	6.2 / 34	0.1 / 0.3	0 / 0.2	0 / 0.4	0 / 0.3
ss_so	7.7 / 38.2	6 / 33.6	4.2 / 33	6.9 / 36.2	0.1 / 6	0.3 / 11.1	0.1 / 9.9	0.1 / 8.8
ss_ti	2.6 / 18.3	0.8 / 11	0.4 / 10	1.5 / 15	0 / 0.3	0 / 0.3	0 / 0.2	0 / 0.1
ss_yo	3.6 / 21.7	1.8 / 16.7	1.2 / 15.9	2.2 / 18.3	0.1 / 4.2	0 / 3.1	0.1 / 7.6	0 / 3.4
ss_zh	19.9 / 20	4.1 / 17.8	4 / 19.3	4.2 / 19.5	0.1 / 0.7	2.5 / 5.3	3.4 / 8.3	3 / 7.9
ti_ckb	5.7 / 39.5	3.4 / 33	1.3 / 26.9	4.4 / 36.3	0 / 0.2	0 / 1.9	0 / 0.3	0 / 0.1
ti_cs	14.5 / 40.9	9.1 / 32.8	11.1 / 34.6	11.8 / 37.2	0.1 / 1.7	0.8 / 11.5	2.4 / 23.2	2.9 / 25.6

ti_cy	18.5 / 45.8	12.5 / 34.9	15.5 / 37.2	16.1 / 40.1	0 / 1.8	0.3 / 14.4	0.1 / 5.6	0.1 / 7.8
ti_en	26.4 / 52.4	15.2 / 41.5	17.2 / 45	20.6 / 47.1	0.6 / 10.2	8 / 29.9	9.4 / 34.6	9.3 / 35
ti_et	12 / 42.8	7.5 / 34.9	7.4 / 34.9	9.3 / 37.6	0.1 / 3.5	0.3 / 10.5	1.6 / 25.8	2.2 / 26.7
ti_eu	8.1 / 41.7	5.8 / 36.1	4.5 / 34.4	7.6 / 40.4	0 / 2.3	0.8 / 24.9	1.3 / 25.2	0.5 / 22.9
ti_fon	1.9 / 20.8	0.8 / 14.2	0.3 / 8.8	0.9 / 14.6	0 / 0.2	0 / 0.3	0 / 1.4	0 / 3
ti_fr	22.1 / 49	15.6 / 39.6	19.5 / 42.6	20.2 / 44.4	0.2 / 3.6	4.7 / 26.2	6.1 / 31.5	3.9 / 31
ti_ln	13 / 44.3	0.5 / 13.5	0.2 / 8.8	0.2 / 0.9	0 / 0.4	0.3 / 10.1	0 / 3.9	0.1 / 6.5
ti_mr	8.2 / 38.6	3.4 / 26.2	1.9 / 23.8	4.6 / 30.3	0 / 1.2	0.7 / 18.8	0.6 / 21.1	0.3 / 18.4
ti_ny	8.4 / 41.2	3.4 / 30.7	1 / 22.7	5.3 / 34.7	0 / 0.4	0 / 0.6	0 / 2.5	0 / 1.1
ti_or	6.7 / 38	4.5 / 31.2	5.1 / 32	5.5 / 34.7	0 / 0.2	0 / 0.2	0 / 0.7	0 / 0.3
ti_so	7.2 / 39.3	4.1 / 32.6	2.3 / 28	5.1 / 35.6	0.1 / 0.4	0.1 / 2.5	0 / 3.5	0 / 3.8
ti_ss	4.8 / 40.5	1.8 / 25.8	0.5 / 17.1	2.5 / 29.7	0.1 / 0.4	0 / 0.9	0 / 0.8	0 / 1.7
ti_yo	4.3 / 24.4	1.4 / 16.7	0.6 / 12.9	1.6 / 17.7	0 / 0.6	0.2 / 7.3	0 / 1.1	0 / 4.9
ti_zh	22.2 / 20.9	2.6 / 16.9	3 / 18.5	4.5 / 21.4	0.3 / 0.7	1.5 / 3.5	3.9 / 5.7	5.2 / 9.4
yo_ckb	5.5 / 36.6	2.4 / 27.5	1.4 / 27.3	3.5 / 31.1	0 / 0.2	0 / 2	0 / 3.2	0 / 2.8
yo_cs	12.7 / 36.6	7.3 / 27.1	9.9 / 31.4	9.7 / 31.7	0.1 / 3.3	0.5 / 18.1	1.6 / 15.1	0.6 / 16.7
yo_cy	16.9 / 42.4	10.3 / 30.1	12.7 / 34.7	13.9 / 35.5	0.1 / 3.7	0.4 / 13.5	0.1 / 9.8	0.1 / 6.9
yo_en	23.9 / 47.9	11.8 / 34.9	12.7 / 40.3	17.8 / 40.7	0.4 / 12.5	5.2 / 23.5	4.2 / 24.3	6.5 / 26.1
yo_et	10.8 / 38.8	5.3 / 29	7.4 / 33.7	8.5 / 34.3	0.1 / 4.5	1.7 / 18.9	1.7 / 20.1	1.3 / 20.8
yo_eu	7.2 / 38.8	3.2 / 29.6	4.9 / 33.9	6.7 / 37	0 / 3.9	0.3 / 19.4	0.4 / 18.2	0.3 / 17.2
yo_fon	1.8 / 19.7	1 / 13.7	0.5 / 11.2	1.1 / 14.9	0 / 0.4	0.1 / 5.2	0 / 4.1	0 / 3.1
yo_fr	19.1 / 44.1	11.5 / 31.5	16.2 / 38	15.9 / 38.4	0.3 / 7	3.2 / 21.8	1.5 / 20.2	1.9 / 22.6
yo_ln	12.4 / 42.8	0.5 / 11.4	0.4 / 11.1	1 / 16.2	0.1 / 4.3	0.3 / 9.1	0.1 / 8	0.1 / 7.7
yo_mr	6.9 / 34.9	1.7 / 20.9	1 / 21.1	3.8 / 27.2	0.1 / 2.5	0.3 / 13.7	0.2 / 12.9	0.1 / 10.3
yo_ny	8.9 / 40.2	2.4 / 25.5	2 / 28.4	6 / 33.9	0.1 / 3.9	0.5 / 9.4	0.1 / 6.2	0 / 5.3
yo_or	5.2 / 33	2.9 / 26.2	2.9 / 28.2	4 / 30.5	0 / 0.3	0 / 0.3	0 / 0.2	0 / 0.2
xx2en	**35.5 / 59.6**	29.7 / 54.4	30.9 / 55.4	31.9 / 56.4	2.0 / 13.3	20.5 / 44.1	22.3 / 46.9	22.4 / 47.6
en2xx	**20.7 / 50.1**	17.3 / 44.1	17.8 / 44.7	18.6 / 45.7	0.4 / 5.7	8.1 / 26.7	8.7 / 29.0	8.7 / 28.8
Mean	**28.2 / 54.9**	23.5 / 49.2	24.4 / 50.0	25.3 / 51.1	1.2 / 9.6	14.3 / 35.5	15.6 / 38.0	15.6 / 38.2
xx2yy	**13.7 / 40.5**	8.8 / 31.2	8.4 / 30.9	10.1 / 34.0	0.3 / 4.1	4.0 / 16.1	4.4 / 17.3	4.2 / 17.1

Table A.11: Evaluation scores on the recently introduced NTREX test set (depicted as <bleu> / <chrf>) for the MT models and language models described in Section 2.1.4.1 and Section 2.1.4.2 compared against unsupervised baselines Baziotis et al. (2023). Note that LM-8B is evaluated on a 50% split of the NTREX data and is not comparable to the MT-model evaluations.

	Baziotis et al. (2023)	**MT-3B**	**MT-7.2B**	**MT-10.7B**	**LM-8B**			
					0-shot	1-shot	5-shot	10-shot
af_en	- / -	57.1 / 75.4	58.5 / 76.5	58.4 / 76.4	7.6 / 30.6	49 / 69.9	49.2 / 69.6	51.8 / 71.9
am_en	- / -	24.3 / 50.3	27.2 / 52.5	27.2 / 52.7	1 / 10.9	13.4 / 37.9	14.6 / 40.4	13.8 / 40.3
ar_en	- / -	36 / 61.9	37.9 / 63.1	37.8 / 63.1	1.1 / 10	23.7 / 51.6	28.6 / 53.9	28.8 / 54.1
az_en	- / -	26.3 / 52.7	29.1 / 54.9	30.1 / 55.8	4.7 / 26.3	15.5 / 40.8	20.6 / 47	21.8 / 48.6

ba_en	- / -	8 / 28.2	9 / 28.4	9.7 / 28.8	0.5 / 7.4	11.1 / 37	9.5 / 34.4	13.2 / 38.7
be_en	- / -	31.5 / 57.6	32.8 / 58.5	35.3 / 59.7	4 / 19.4	18.4 / 40.7	29.1 / 56	30.7 / 56.3
bem_en	- / -	10 / 32.6	12.1 / 33.9	13.3 / 34.7	0.1 / 5.3	4.8 / 20.7	2.7 / 21.7	2.6 / 21.5
bg_en	- / -	38.2 / 63.3	39.3 / 64	39.4 / 64.1	4.3 / 25.4	31.6 / 58.3	31.6 / 58.7	32.6 / 58.3
bn_en	- / -	35.4 / 60.7	37.5 / 61.9	38 / 62.3	0.8 / 6	23.5 / 51.4	22.9 / 50.1	24.4 / 51.9
bo_en	- / -	12.1 / 37.8	13.3 / 38.8	14.5 / 40	0.1 / 2.7	5.9 / 28.4	2.9 / 25.8	4.3 / 27.8
bs_en	- / -	40.7 / 65.1	42.6 / 66.2	41.9 / 66	4.6 / 24.2	27.2 / 52.2	33.6 / 59.4	35.3 / 60.4
ca_en	- / -	41.6 / 66	43.4 / 66.9	43.1 / 66.8	5.4 / 22.9	31.3 / 56.6	34.5 / 61	36.1 / 61.4
ckb_en	- / -	28.3 / 53.4	33.9 / 58	33.5 / 57.9	1.1 / 11.4	7.8 / 27.4	18.8 / 43.6	20.2 / 45.3
cs_en	- / -	39.4 / 64.8	41.1 / 65.7	41.4 / 65.8	5.8 / 29.1	26 / 51.5	33.5 / 59.3	34.1 / 59.7
cy_en	- / -	43.2 / 66.6	45 / 68.1	45.2 / 68.2	3.9 / 24.2	35.2 / 58.8	37.5 / 61.1	37.2 / 60.7
da_en	- / -	42.5 / 65	43.7 / 65.8	43.7 / 65.8	6.3 / 28.6	36.1 / 60.2	37.6 / 61.2	37.5 / 61.2
de_en	- / -	38.9 / 64.6	39.8 / 65.4	40.2 / 65.5	4.3 / 26.4	31.6 / 59.4	33 / 59.1	34 / 60.4
dv_en	- / -	22.2 / 48.8	24.6 / 50.5	25 / 50.8	0.6 / 6.8	7.9 / 29.6	11.1 / 36.9	12.6 / 39.5
dz_en	- / -	6.4 / 28.2	6.4 / 27.9	6.7 / 28.7	0 / 1.6	0.9 / 13.9	1.1 / 17.5	0.9 / 15
ee_en	- / -	13.2 / 35.1	16.3 / 38.4	16.7 / 38.5	0.3 / 7.9	3.2 / 22.1	1 / 15.2	2.8 / 19.6
el_en	- / -	43 / 65.8	43.3 / 66.1	44.6 / 66.7	5.2 / 23.8	33.9 / 58.7	31.7 / 58	32.3 / 59
en_GB_en	- / -	89.3 / 94.4	90.3 / 94.6	90.5 / 94.9	7.9 / 26	97.5 / 98.7	99.4 / 99.8	99.4 / 99.7
en_IN_en	- / -	87.1 / 93.6	86.4 / 93	86.9 / 93.8	10.1 / 27	35.1 / 45.3	93.7 / 98.3	95.6 / 98.7
en_af	- / -	42.3 / 66.7	44.5 / 68.4	44.6 / 68.3	2.9 / 16.5	35 / 61.4	37.7 / 63.3	37.8 / 63.4
en_am	- / -	3.8 / 23	4.5 / 23.8	4.2 / 24.3	0.1 / 0.7	0.2 / 0.8	0 / 0.6	0 / 0.4
en_ar	- / -	24.7 / 54.2	24.6 / 54.3	24.9 / 54.7	0.4 / 2.9	6.7 / 38.1	15.4 / 45.3	12.1 / 44
en_az	- / -	14.6 / 46.7	15.2 / 47.5	15.5 / 48.2	0.4 / 8.6	10.4 / 41.6	10.6 / 42.9	7.4 / 40.9
en_ba	- / -	3.3 / 24.3	4 / 26.3	5.2 / 28.9	0 / 0.4	0.3 / 13.4	0.1 / 9.8	0.1 / 10.5
en_be	- / -	21.9 / 44.8	24.2 / 47.3	26 / 50.2	0.7 / 6	20.6 / 49.4	21.6 / 50.3	21.9 / 50.6
en_bem	- / -	2.4 / 4	3.9 / 19.5	2.1 / 3.4	0.1 / 6.7	0.1 / 8.7	0.2 / 10.6	0.1 / 5
en_bg	- / -	34.1 / 59.5	34.8 / 59.9	34.6 / 59.7	1.4 / 8.8	21.1 / 48.7	27.3 / 53.4	29.1 / 55.1
en_bn	- / -	13.8 / 48.5	14 / 49.2	14.6 / 49.8	0.4 / 7.4	8.6 / 40.8	7.4 / 41.5	5.8 / 42
en_bo	- / -	12 / 25.2	14.7 / 29.1	14 / 28.6	0 / 0.9	0 / 1.8	0 / 1.6	0 / 1.1
en_bs	- / -	29.4 / 58.6	30.3 / 59.4	30.2 / 59.3	0.7 / 10.5	23.6 / 53	24.9 / 54.2	25.3 / 54.3
en_ca	- / -	39.9 / 63	40.9 / 63.7	40.5 / 63.4	2.3 / 15.1	30.1 / 55.8	31.6 / 57.7	32.1 / 58.1
en_ckb	- / -	8.5 / 42.4	9.2 / 42.5	9.2 / 42.2	0 / 0.2	0.1 / 10.7	0.2 / 10.9	0.1 / 9.3
en_cs	- / -	32.4 / 59.4	32.9 / 59.6	32.6 / 59.7	1 / 10.8	22.2 / 50.2	22.7 / 51.2	20.9 / 51.5
en_cy	- / -	29.6 / 56.4	33 / 59.7	33.3 / 59.8	0.3 / 9.3	3.4 / 25.2	2.6 / 24.2	3 / 29.1
en_da	- / -	40.4 / 63.9	40.9 / 64.3	40.7 / 64.1	2 / 13.5	34.8 / 59.2	35.5 / 59.9	36.6 / 60.6
en_de	- / -	34.5 / 61.8	35.2 / 62.1	34.7 / 61.8	2 / 16.2	21.3 / 49.9	23.5 / 53.1	22.1 / 52.7
en_dv	- / -	3.5 / 40.7	4.7 / 44.4	4.7 / 44.7	0 / 0.2	0.3 / 0.4	0.2 / 0.4	0.1 / 0.4
en_dz	- / -	22 / 35.8	22 / 35.5	22.7 / 37.2	0 / 0.1	0 / 2.1	0 / 0.3	0 / 1.1
en_ee	- / -	6.1 / 29.1	7.4 / 31.6	7.2 / 30.5	0.2 / 6.2	0.4 / 13.5	0.2 / 6.9	0.1 / 4.5
en_el	- / -	37.4 / 61.5	37.4 / 61.6	37.6 / 61.6	0.9 / 3.7	26.7 / 50.8	28.2 / 52.4	28 / 52.6
en_en_GB	- / -	6.5 / 9.1	34.4 / 48.4	5.8 / 7.1	8.3 / 21.8	99.1 / 99.7	99.4 / 99.8	99.5 / 99.8
en_en_IN	- / -	5.7 / 8.7	21.2 / 28	4.1 / 5.5	4.4 / 15.4	91.2 / 97.4	92.3 / 98	92.1 / 97.9
en_es	- / -	42 / 65.4	42.3 / 65.6	42.2 / 65.5	1.4 / 15.4	28.6 / 53.7	34.1 / 59	33.5 / 58.5
en_es_MX	- / -	2.4 / 5.7	7.6 / 28.1	1.7 / 4	0.8 / 9.7	34.1 / 58.8	36.9 / 60.7	36.5 / 60.9
en_et	- / -	26.4 / 59	26.7 / 59.1	27.3 / 59.4	0.5 / 10.2	11.2 / 42.8	20.2 / 52.9	21.8 / 53.7
en_eu	- / -	14.8 / 51.3	15 / 51.2	15.4 / 52.3	0.2 / 9.4	9.9 / 45.6	9.8 / 45.7	7.4 / 45.7
en_fa	- / -	19.1 / 45	19.5 / 44.9	19.5 / 44.9	0.5 / 3.1	14.3 / 40.7	14.9 / 42.4	14.8 / 43
en_fa_AF	- / -	0.3 / 0.9	0.2 / 0.8	0.2 / 0.8	0 / 0.2	12.4 / 39.6	13.5 / 41.1	11.4 / 40.1
en_ff	- / -	4.8 / 27.6	5.3 / 28.1	5.1 / 28.8	0.2 / 5.5	0.1 / 7.9	0.1 / 5.5	0.1 / 4.3
en_fi	- / -	21.5 / 57.1	21.7 / 57.4	22 / 57.3	0.5 / 11	14.5 / 47.5	16.4 / 50.5	17.2 / 50.4

en_fil	- / -	25.3 / 55.4	25.8 / 55.6	25.3 / 55.4	1.6 / 14.6	26.5 / 54	30 / 57.6	28.6 / 55.8
en_fj	- / -	6.4 / 27.5	7.3 / 28.6	7.7 / 29	0.3 / 6.8	0 / 4.6	0.2 / 7.3	0.1 / 5
en_fo	- / -	16.4 / 37.9	18.6 / 40.8	16.6 / 38.2	0.4 / 9	5.6 / 29.1	7.5 / 32.5	7.2 / 32.7
en_fr	- / -	38.3 / 62.9	39.2 / 63.5	39.2 / 63.5	0.8 / 12.7	28.8 / 54.8	30.2 / 56.2	26.7 / 56.6
en_fr_CA	- / -	2.1 / 5.7	4.2 / 27.8	2.5 / 7.6	0.6 / 8.6	29.6 / 55.8	29.7 / 56.5	30 / 57.6
en_ga	- / -	26.8 / 54.9	27.5 / 55.5	27.1 / 55.2	0.4 / 10.2	1.2 / 16.2	0.9 / 15.4	0.8 / 15.9
en_gl	- / -	39.9 / 64.5	40.6 / 65	40.5 / 64.8	1.3 / 13.2	31.3 / 57.5	32.6 / 58.6	30.6 / 58.1
en_gu	- / -	14.9 / 47.5	15.3 / 48	15.4 / 48.2	0.1 / 0.3	0.4 / 1.2	0.1 / 1.7	0.1 / 1.3
en_ha	- / -	21 / 49	21.8 / 50.1	22.8 / 51.1	0.2 / 4.2	0.4 / 11.4	0.3 / 9.9	0.3 / 9
en_hi	- / -	25.5 / 50.2	25.7 / 50.2	25.9 / 50.6	0.9 / 7.8	16.7 / 42.9	15.6 / 42.8	12.6 / 42.8
en_hmn	- / -	16.6 / 42.1	17.5 / 42.9	17.7 / 43.3	0.2 / 6.1	1.1 / 12.4	0.6 / 10.3	0.5 / 9.4
en_hr	- / -	34.9 / 62	35.5 / 62.7	35.6 / 62.8	3.7 / 21.3	25.7 / 54	26.7 / 55.3	25.1 / 55.3
en_hu	- / -	21 / 50.8	21.4 / 51.2	21.7 / 51.4	0.6 / 10.8	10.6 / 36.9	11.1 / 42.3	12.3 / 42.5
en_hy	- / -	15.7 / 46.5	18.5 / 51	19.1 / 51.7	0.1 / 0.5	0.4 / 0.8	0.5 / 1	0.5 / 1.2
en_id	- / -	38.7 / 65.2	39.6 / 65.8	39.8 / 65.8	0.7 / 9	32.8 / 60.5	35 / 62.7	34.8 / 62.7
en_ig	- / -	13.5 / 37.9	16.4 / 43.2	16.8 / 42	0.2 / 5.8	0.2 / 9.7	0.1 / 5.3	0.2 / 5.5
en_is	- / -	25.1 / 52.8	25.3 / 53	25.3 / 53.3	2.6 / 16.8	22.5 / 48.9	26.3 / 52.3	26 / 51.9
en_it	- / -	38.8 / 64	39.2 / 64.1	39 / 64.1	1.2 / 14.7	28.7 / 55.8	29.7 / 56.7	30.2 / 57
en_iw	- / -	24.3 / 53	22.6 / 48.5	25.4 / 54.2	0.5 / 2.6	14 / 42.3	15.9 / 44.6	15.7 / 44.8
en_ja	- / -	6.3 / 33.6	6.8 / 34.3	5.9 / 33.9	0 / 3.2	0.1 / 26.5	0.1 / 26.5	0.2 / 24.8
en_ka	- / -	2.7 / 15.3	1.6 / 11.9	3.1 / 15.5	0.6 / 6.4	10.2 / 42.1	11.5 / 44.8	9.7 / 46
en_kk	- / -	12.3 / 45.5	13.2 / 47.1	13.7 / 47.3	0.3 / 5.9	8.4 / 41.2	7.7 / 40.8	6.7 / 41.2
en_km	- / -	6 / 40.8	6 / 41.3	4.8 / 43.6	0.1 / 1.1	1.6 / 27.4	0.7 / 32.3	0.3 / 29.3
en_kn	- / -	10.8 / 48.4	12.1 / 50	11.9 / 49.8	0.1 / 0.5	0.2 / 4.1	0.1 / 4.9	0.1 / 4.9
en_ko	- / -	10.9 / 31.7	11.6 / 32.4	11.9 / 32.7	0.2 / 2.8	5.6 / 22.1	3.5 / 22.8	2.9 / 22.8
en_ku	- / -	0.2 / 0.5	0.2 / 0.6	0.3 / 0.8	0.1 / 6.4	0.6 / 15.8	0.5 / 16.3	0.3 / 14.3
en_ky	- / -	9 / 42.1	10.6 / 45	10.4 / 44.2	0.1 / 1.1	1.3 / 22.7	0.9 / 24.5	0.8 / 24.9
en_lb	- / -	8.5 / 29.5	8.9 / 29.8	9.6 / 31	0.4 / 10	3.5 / 32.4	4.1 / 32.7	4.6 / 33.9
en_lo	- / -	11.1 / 38	11.1 / 38.4	11.7 / 39.1	0.2 / 1.4	1.3 / 19.4	2.7 / 31.6	1.9 / 32.6
en_lt	- / -	24.3 / 56.1	26 / 56.7	25 / 56.3	0.6 / 11.8	15.8 / 46.8	17.8 / 50	17.4 / 50.3
en_lv	- / -	22.6 / 54.1	22.9 / 54.1	23.4 / 54.6	1 / 10.2	17.6 / 48.1	18.8 / 49.4	19.1 / 50.6
en_mey	- / -	0.1 / 0.6	0.1 / 0.5	0.1 / 0.6	0 / 0.2	2.9 / 29.9	0.7 / 21	0.8 / 23.2
en_mg	- / -	16.8 / 48.5	17.7 / 50.1	17.6 / 49.6	0.4 / 6.7	1.3 / 19.4	0.6 / 17.1	0.3 / 14
en_mi	- / -	20.9 / 46.4	19.7 / 45.8	21.5 / 47.1	0.5 / 7.9	0.8 / 13.6	0.4 / 11.2	0.4 / 10
en_mk	- / -	37 / 63.9	37.6 / 64.1	37.6 / 64.1	1 / 4	27.2 / 55.3	28.9 / 57.1	30.4 / 58.5
en_ml	- / -	7.1 / 44.2	8 / 44.4	8.3 / 45.3	0.3 / 8.8	5.4 / 37	4.6 / 39.7	2.1 / 37.9
en_mn	- / -	10.7 / 41.6	12 / 44	12.4 / 45	0.1 / 0.5	0.6 / 1.6	0.4 / 3.5	0.4 / 4.1
en_mr	- / -	7.5 / 36	7.3 / 34.2	8.4 / 37.3	1.2 / 14.1	7.8 / 39.6	7.4 / 40	6.8 / 40.3
en_ms	- / -	33.6 / 61.8	35.1 / 62.8	35.3 / 62.9	1.8 / 14.2	29.1 / 56.6	31.6 / 62	33 / 62.2
en_mt	- / -	45.7 / 66.6	47.1 / 67.3	47.6 / 67.5	0.4 / 10	3.7 / 25.6	3.7 / 34.2	3.3 / 33.7
en_my	- / -	1.7 / 16.7	1.5 / 16.2	1.4 / 16.1	0 / 0.5	0 / 5.4	0 / 5.1	0 / 6.4
en_nd	- / -	1.3 / 3	1.8 / 18.2	1 / 2.2	0.1 / 2.5	0.3 / 20.2	0.2 / 13.3	0.2 / 13
en_ne	- / -	7.5 / 35.9	7.9 / 37	8.1 / 37.8	0.1 / 0.5	1.4 / 23.9	1.7 / 30.3	1 / 26.5
en_nl	- / -	36.7 / 62.9	37.5 / 63.4	37.4 / 63.3	2 / 15.7	25 / 51.7	28.8 / 56.6	29.8 / 57.8
en_nn	- / -	34.6 / 60.9	35.8 / 61.6	35.8 / 61.4	4.5 / 14.4	24.9 / 53.2	29.9 / 56.1	30.6 / 57
en_no	- / -	38.7 / 63.6	39 / 63.7	38.9 / 63.6	1.9 / 12.7	33.1 / 59	34.9 / 60.3	34.8 / 60.3
en_nso	- / -	7.9 / 31.2	9.1 / 33.5	9.2 / 33.7	0.3 / 4.6	0.9 / 13.8	0.5 / 13.5	0.4 / 12.4
en_ny	- / -	12 / 41.6	14.8 / 45.5	14.4 / 44.6	0.7 / 8	1.1 / 18.3	0.7 / 17.8	0.7 / 18.3
en_om	- / -	1.1 / 23.2	1.8 / 30.4	2 / 30.3	0.1 / 6.5	0.1 / 9.7	0.1 / 7.3	0.1 / 7.6
en_pa	- / -	18.2 / 44.1	18.9 / 45.2	19.2 / 45.6	0.1 / 0.3	1 / 9	0.6 / 2.6	0.2 / 1.4

en_pl	- / -	27.6 / 55.5	28.1 / 55.8	28.2 / 55.9	1.4 / 12.2	18.1 / 45.3	15.2 / 46.7	14.9 / 45.4
en_ps	- / -	10.5 / 33.6	10.8 / 34.3	10.8 / 34.4	0.1 / 0.5	0.5 / 10.3	0.3 / 8.6	0.3 / 7.6
en_pt	- / -	41.7 / 64.9	42.3 / 65.3	42.3 / 65.3	3.3 / 18.3	30 / 55.8	33.4 / 58.8	34 / 59.7
en_pt_PT	- / -	2.7 / 5.9	8.9 / 34.5	3 / 18	2.9 / 13.4	23.9 / 49.3	31.7 / 57.4	32.7 / 57.9
en_ro	- / -	34.5 / 59	35.3 / 59.4	35 / 59.2	4.4 / 19.8	28.1 / 52.8	30.2 / 55	30.8 / 55.4
en_ru	- / -	30.7 / 55.8	31.9 / 56.7	31.8 / 56.7	1 / 2.9	18 / 41.8	22.3 / 48.5	23 / 48.9
en_rw	- / -	3.2 / 18.6	0.3 / 8.3	0.9 / 12	1 / 9.5	1.1 / 18	0.8 / 17.5	0.5 / 16.5
en_sd	- / -	10 / 35.8	11.4 / 37.1	11.2 / 37.4	0.1 / 0.6	0.3 / 8.7	0.2 / 8.8	0.3 / 10.9
en_shi	- / -	0.3 / 0.7	0.2 / 0.7	0.2 / 0.7	0 / 0.3	0.1 / 4.3	0 / 3.4	0 / 3.3
en_si	- / -	11 / 42.6	10.9 / 42.2	11.4 / 44.2	0.1 / 0.5	0.4 / 1.1	0.6 / 1	0.4 / 1.2
en_sk	- / -	33.9 / 60.3	34.2 / 60.4	34.3 / 60.7	2 / 14.3	26.4 / 53.1	25.6 / 53.2	27.5 / 54.1
en_sl	- / -	32.4 / 59.6	33.2 / 60	33.4 / 60.1	1.1 / 12.4	21.1 / 50.5	23.9 / 52.5	24.9 / 53
en_sm	- / -	23.4 / 48.7	25.4 / 51.2	27 / 51.9	0.4 / 7.5	1 / 16.6	0.5 / 12.6	0.4 / 11.8
en_sn	- / -	8.9 / 38.5	10.6 / 41.4	11.4 / 43.5	0.3 / 6	0.5 / 15.3	0.5 / 16.8	0.5 / 17
en_so	- / -	14.2 / 46.9	14.5 / 48	14.8 / 48.1	0.3 / 7.6	2.1 / 14.7	0.3 / 12.5	0.4 / 14.3
en_sq	- / -	31.8 / 57.9	32.3 / 58.3	31.9 / 58	1.9 / 11.8	30.4 / 56.3	31.4 / 56.8	31.9 / 57
en_sr	- / -	21.3 / 45	21.7 / 45.4	21.8 / 45.5	0.2 / 1.3	4.2 / 21.3	14.9 / 39.3	9.3 / 34.3
en_sr_Latn	- / -	0.9 / 3.6	0.9 / 19	0.8 / 2.5	1.1 / 5	21.4 / 50.1	22.9 / 51.1	23 / 51.5
en_ss	- / -	5.7 / 35.4	6.6 / 36.7	6.9 / 38	0.2 / 6.7	0.1 / 6.1	0.2 / 12.9	0.2 / 11.4
en_sv	- / -	43.3 / 67.5	43.7 / 67.6	43.2 / 67.3	2 / 13.7	35 / 60.4	36 / 62.5	35.8 / 61.8
en_sw	- / -	31.7 / 57.2	30.9 / 56.2	29 / 54.7	0.8 / 9.5	2.4 / 23	1.7 / 25	2.8 / 31.2
en_ta	- / -	7.8 / 45.1	8.3 / 45.3	8.4 / 45.8	0.4 / 11.7	2.1 / 34.1	4.6 / 40.3	2.6 / 40.1
en_te	- / -	9.2 / 44.3	9.8 / 45.1	10 / 45	0.5 / 9.6	5.4 / 35.8	5.4 / 39.8	3.6 / 38.5
en_tg	- / -	3.2 / 21	5.5 / 24.4	3.4 / 19.5	0.1 / 0.5	0.8 / 15.3	0.8 / 19.7	0.7 / 20.6
en_th	- / -	6.6 / 40.9	5.6 / 40.5	5.8 / 39.4	0.2 / 10.1	11.8 / 47.4	9 / 48.3	12.6 / 49.1
en_ti	- / -	2.3 / 15.7	3.3 / 19.6	3.9 / 20.4	0.2 / 0.5	0 / 1.1	0 / 0.9	0 / 1.1
en_tk	- / -	12.2 / 42	14.2 / 46.2	13.8 / 45.6	0.2 / 6.7	1.1 / 18.3	0.5 / 19.2	0.6 / 19.4
en_tn	- / -	16.3 / 36.9	19.9 / 41.5	19.1 / 40.9	0.5 / 7.3	0.3 / 9.1	0.3 / 10.5	0.4 / 11.1
en_to	- / -	3.9 / 25.5	3.3 / 26.4	4.9 / 27.8	0.1 / 3.6	0.3 / 11.2	0.2 / 9.3	0.2 / 8.3
en_tr	- / -	25 / 55.1	25.4 / 55.4	25.9 / 55.7	0.8 / 11.6	11.6 / 43.4	11.2 / 44.4	10.4 / 44.7
en_tt	- / -	5 / 27.3	8 / 32.4	9.7 / 36	0.1 / 0.5	0.7 / 19.7	0.7 / 22.2	0.4 / 20.7
en_ty	- / -	0.6 / 2.4	1.3 / 12.8	0.4 / 2.1	0.4 / 9.2	0.6 / 14.1	0.4 / 12.6	0.4 / 11.3
en_ug	- / -	4.7 / 29.9	5.4 / 31.3	5.8 / 31.5	0.1 / 0.3	0.1 / 1.4	0.2 / 2.3	0 / 1.2
en_uk	- / -	24.3 / 51.4	25.5 / 52.6	25.6 / 52.5	1.4 / 5.6	18.5 / 45.9	20.4 / 48.2	18.1 / 46.9
en_ur	- / -	2.3 / 16.3	2.3 / 16.3	2.4 / 16.4	0.2 / 1.4	7.1 / 32	4.3 / 31.5	2.5 / 28.2
en_uz	- / -	0.7 / 1.5	0.7 / 1.4	0.6 / 1.3	0.1 / 7.3	1.3 / 26.8	1.8 / 31.9	2 / 33.9
en_ve	- / -	7 / 31.1	7.8 / 32	8.5 / 32.9	0.2 / 4.2	0.3 / 11.9	0.1 / 8.1	0.1 / 8.1
en_vi	- / -	0.1 / 12.5	0.1 / 12.4	0.1 / 12.4	1.4 / 9.2	34.4 / 53.4	34.7 / 53.8	34.6 / 54
en_wo	- / -	0.4 / 9.8	0.5 / 10	0.5 / 12.1	0.3 / 6.4	0.4 / 11.4	0.2 / 8.9	0 / 5.3
en_xh	- / -	10.3 / 45.1	10.9 / 47.1	11.3 / 47.5	0.2 / 6.8	0.3 / 14	0.2 / 12.5	0.2 / 13.7
en_yo	- / -	5.4 / 20.9	5.5 / 21.6	5.7 / 21.2	0.2 / 4.2	0.1 / 4.6	0 / 6.1	0.2 / 10.5
en_yue	- / -	0 / 1.3	0 / 1.2	0 / 1.2	0 / 1.9	0.1 / 13.2	0 / 14.6	0 / 15.7
en_zh	- / -	3.8 / 29.7	3.9 / 30.1	3.9 / 30.1	1.9 / 2.7	19.5 / 21.2	23.6 / 21.3	23.1 / 23.2
en_zh_Hant	- / -	1 / 8.2	0.9 / 8.8	0.7 / 5.3	0.4 / 4	6 / 23.1	6.5 / 22.9	3.8 / 22.7
en_zu	- / -	2.7 / 24.6	2.6 / 25.2	2.6 / 24.7	0.5 / 7.3	0.5 / 15.2	0.2 / 13.9	0.2 / 12.6
es_MX_en	- / -	50.4 / 71.7	51.9 / 72.7	52.3 / 72.7	6.1 / 24	40.5 / 64.9	41.3 / 65.4	41.1 / 65.3
es_en	- / -	41.1 / 66.4	42.4 / 67.1	42.9 / 67.3	5.9 / 26.2	33.1 / 60	35.2 / 61.9	35.8 / 62
et_en	- / -	35.1 / 61.8	36.8 / 62.8	36.8 / 62.8	8.6 / 31.2	28.2 / 55.8	30.3 / 56.7	30.3 / 56.6
eu_en	- / -	31.8 / 56.8	34 / 58.3	34 / 58.2	5.5 / 25.7	22.4 / 48.8	23.4 / 49.6	23.8 / 50.1
fa_AF_en	- / -	32.3 / 58.6	33 / 59	34.1 / 59.8	2.4 / 14.8	18.2 / 42.1	24.6 / 51.1	25.6 / 52

fa_en	- / -	32.2 / 58.6	34 / 59.7	33.7 / 59.5	3.7 / 18.7	16.3 / 37.6	24.1 / 50.2	25.4 / 51.6
ff_en	- / -	14.4 / 35.8	14.2 / 36	15.8 / 37.5	0.3 / 9.6	3.3 / 20.5	4 / 18.7	2.4 / 18.8
fi_en	- / -	30.3 / 57.8	31.7 / 58.8	31.5 / 58.6	4.2 / 24.6	24.3 / 52.5	26.4 / 52.8	28.1 / 55.1
fil_en	- / -	45.7 / 66.9	48.2 / 68.5	48.1 / 68.7	4 / 19.3	38.3 / 61.7	36.2 / 59.5	39.9 / 62.6
fj_en	- / -	6.6 / 26.6	8.4 / 28.5	8.2 / 28.4	0.5 / 10.3	4.7 / 18.7	6.2 / 27.1	3.7 / 21.7
fo_en	- / -	32.9 / 56.7	37 / 59.8	38.6 / 61	8 / 28.7	28.6 / 52.8	33.2 / 57.1	33.6 / 56.9
fr_CA_en	- / -	43.2 / 66.8	44.1 / 67.4	43.8 / 67.4	5 / 25.8	26.6 / 56	37.3 / 62.8	36.7 / 62.5
fr_en	- / -	38.8 / 63.5	39.5 / 64.1	39.2 / 63.9	4.5 / 26.1	31.1 / 57.6	32.6 / 57.8	32.6 / 58.8
ga_en	- / -	38.7 / 64.1	41 / 65.7	41.6 / 65.7	4.2 / 23	24.6 / 48.7	29.2 / 53.5	31.5 / 55.3
gl_en	- / -	43.7 / 66.8	45.3 / 67.9	45.1 / 67.8	6 / 19.5	33.9 / 59.7	35.3 / 60.3	36.9 / 62.3
gu_en	- / -	34.9 / 62	38.3 / 64	38 / 64.1	2.2 / 16.6	21.1 / 48	24.2 / 52.1	25.4 / 53.6
ha_en	- / -	35 / 56.9	39.2 / 59.9	38.7 / 60	1.1 / 14.7	12.5 / 33	21 / 43.6	18.9 / 43.3
hi_en	- / -	35.7 / 61.8	38.7 / 63.4	38.2 / 63.1	2.7 / 14.7	20 / 44.9	24.6 / 52.3	26.7 / 54.3
hmn_en	- / -	23.4 / 47.5	25.5 / 49.6	27 / 50.2	1 / 12.3	10.8 / 30.5	13.4 / 38.2	14.3 / 37.7
hr_en	- / -	41.3 / 65.5	42.4 / 66.5	42.3 / 66.4	6.6 / 26.9	29.9 / 55.6	31.7 / 57.6	35.2 / 60.6
hu_en	- / -	27.8 / 55.6	27.6 / 55.8	28.1 / 55.8	3.5 / 25.6	20.1 / 47	21.6 / 49	21.3 / 49.2
hy_en	- / -	30.7 / 55.8	34.8 / 59.3	34.1 / 58.9	3.2 / 18.7	17.8 / 44.4	22.9 / 48.8	23.7 / 50.7
id_en	- / -	43 / 66.4	45.1 / 67.8	45 / 67.8	3.8 / 18.4	34.8 / 59.9	37.3 / 61.5	38 / 61.9
ig_en	- / -	27.5 / 51.1	31.9 / 55.3	32.7 / 55.8	1.1 / 14.4	9.4 / 28.1	11.4 / 33.4	11.8 / 33.1
is_en	- / -	35.3 / 60.5	37.3 / 61.8	37.7 / 61.8	9.1 / 31.5	29 / 54.9	30.5 / 55.3	33 / 57.9
it_en	- / -	42.3 / 66.2	43.2 / 66.8	43.2 / 66.9	7.5 / 28.6	36.8 / 61.5	36 / 60.5	38.5 / 62.4
iw_en	- / -	40.5 / 63.8	42.9 / 65.7	42.9 / 65.6	4.3 / 19.1	29.2 / 55.7	31 / 56.9	31.1 / 57.4
ja_en	- / -	27.1 / 54.5	28.7 / 55.1	29.1 / 55.5	0.3 / 1.2	15 / 40.9	19.2 / 46.6	18.7 / 45.9
ka_en	- / -	28.8 / 53.9	31.8 / 56.1	32 / 56.3	2.6 / 14.1	20.1 / 45	26.3 / 52.4	25.3 / 52.6
kk_en	- / -	24.1 / 52.1	26.4 / 53.6	25.7 / 53.3	3.7 / 18.6	16.4 / 42.9	18.3 / 44.3	18 / 44
km_en	- / -	33.6 / 59.6	37.8 / 61.4	37.3 / 61.7	1.9 / 16.1	21.1 / 47.2	26.4 / 52.2	27.2 / 53.5
kn_en	- / -	29.2 / 56.8	31.4 / 57.9	31.1 / 58.3	1.3 / 11.5	13.7 / 38.9	18.8 / 45.8	20.4 / 48.2
ko_en	- / -	30.9 / 57.1	33.1 / 58.6	33.2 / 58.8	1.1 / 4.5	19 / 45.8	18.7 / 45.2	19 / 47
ku_en	- / -	27.6 / 52.5	30.5 / 54.4	30.7 / 54.9	2.9 / 18.3	17.2 / 41.7	16.5 / 43.2	20.2 / 44.2
ky_en	- / -	23.8 / 50.5	27.5 / 53.6	26.2 / 52.6	2.6 / 16.1	15.4 / 41.5	15.9 / 41.1	17 / 43.2
lb_en	- / -	21.9 / 49.8	24.4 / 51.6	26 / 52	4.2 / 23.8	29.5 / 53.3	32.3 / 56.6	30.6 / 56.1
lo_en	- / -	34.9 / 58.2	37.5 / 59.7	37.9 / 60.9	1.2 / 11.8	23.4 / 49	28.5 / 53.9	28.1 / 53.5
lt_en	- / -	32 / 59.6	33.9 / 60.6	34.1 / 60.6	4.9 / 25.7	24.3 / 52.3	25.7 / 52.9	26.4 / 52.9
lv_en	- / -	34.6 / 61.2	36.2 / 62.3	36.6 / 62.3	3.4 / 22.7	26.1 / 53.5	29.5 / 55.4	30.6 / 57.1
mey_en	- / -	18 / 45.5	19.7 / 47	20.2 / 46.9	0.3 / 4.2	4.5 / 22.6	11.8 / 38.8	10.4 / 38.8
mg_en	- / -	25.8 / 50.8	28.3 / 52.7	28.7 / 52.9	1.7 / 18.6	14 / 37.6	14.8 / 39.1	16 / 39.5
mi_en	- / -	22.1 / 45.3	20.5 / 44.1	22.5 / 46.3	2.2 / 19.2	10.4 / 35.6	10.8 / 35.7	17.7 / 41.1
mk_en	- / -	44.6 / 67.4	46.3 / 68.5	46.4 / 68.4	8.5 / 28.5	29.3 / 55.2	36.9 / 61.3	37.7 / 62.6
ml_en	- / -	30.3 / 56.8	31.4 / 57.1	32.4 / 58.2	0.8 / 5.7	11.5 / 34.3	17.8 / 45.8	18.8 / 46.5
mn_en	- / -	21.8 / 48.4	24 / 50.4	24.1 / 50.2	1.3 / 10.1	12.9 / 38.5	13.6 / 38.4	14.3 / 40.8
mr_en	- / -	32.5 / 59.3	34 / 60.2	34.9 / 60.9	1.7 / 9.8	12.7 / 35.2	22.6 / 50.7	23.4 / 50.4
ms_en	- / -	40.4 / 63.7	42.7 / 65.4	42.7 / 65.3	3.9 / 20.7	31.3 / 55.8	36 / 60.1	36.4 / 60.3
mt_en	- / -	50.9 / 72.9	52.4 / 73.9	53.5 / 74.3	5.6 / 25	40.3 / 63.2	40.1 / 62.5	41.5 / 65.1
my_en	- / -	14.3 / 40.3	15.7 / 40.9	18.1 / 44.5	0.2 / 5.7	2.8 / 20.2	7.5 / 32.2	9.7 / 34.2
nd_en	- / -	18.8 / 43.1	21.6 / 45.6	21.8 / 45.8	0.2 / 8	8 / 27.2	9 / 32.6	9.1 / 32.1
ne_en	- / -	35.9 / 62	38.1 / 63.3	38.6 / 63.9	1.4 / 9	15.6 / 40.3	22.3 / 48.8	23.5 / 50.9
nl_en	- / -	41.2 / 66.1	42.5 / 66.7	42.5 / 66.6	8.3 / 31.4	35.1 / 60.3	36.5 / 61.4	36.6 / 62.3
nn_en	- / -	46 / 67.3	47.1 / 68.1	47 / 68.1	5.4 / 26.9	39 / 63.2	40.5 / 63.9	41 / 63.3
no_en	- / -	44 / 67	45.1 / 67.7	45 / 67.7	6.1 / 25.6	38.8 / 61.7	38.8 / 63.5	39.7 / 63.5
nso_en	- / -	23.6 / 47.6	27.5 / 51.3	28.1 / 51.7	0.6 / 10.9	3.7 / 26.1	10.1 / 31.4	11.1 / 34.9

ny_en	- / -	28.1 / 49.5	31.6 / 52.9	31.9 / 53.2	1.1 / 13	15.3 / 36.4	17.6 / 38.8	15.4 / 39.4
om_en	- / -	8.4 / 29.2	11.2 / 33.6	11.2 / 33.6	0.5 / 9.8	1.9 / 21.2	4.6 / 24.3	5.2 / 23.2
pa_en	- / -	34.6 / 60.8	38.5 / 62.9	38.2 / 63.1	2.4 / 16.1	16.1 / 42.3	24 / 51	25.4 / 51.8
pl_en	- / -	31.9 / 58.9	32.6 / 59.2	33.3 / 59.7	9.3 / 34.9	24.4 / 52.1	27.4 / 54.4	27.3 / 53.8
ps_en	- / -	22.6 / 49.5	23.9 / 50.5	24.5 / 50.8	1.9 / 14.2	14.1 / 38.8	15.6 / 40.3	16.3 / 42.1
pt_PT_en	- / -	44.7 / 67.4	45.7 / 68.2	45.9 / 68.2	5.2 / 23.6	35.2 / 60.3	36 / 60.8	36.8 / 61.4
pt_en	- / -	43.1 / 65.7	43.9 / 66.2	44.1 / 66.2	6.4 / 24.9	34.4 / 59.3	35.8 / 60.8	36.8 / 61.1
ro_en	- / -	37.7 / 63.9	38.5 / 64.6	39 / 64.7	6.1 / 28.5	22.8 / 46.9	32.3 / 58.6	33.4 / 59.5
ru_en	- / -	30.8 / 57.9	32.3 / 58.8	32.1 / 58.7	3.7 / 19.2	24 / 50.2	26.1 / 53.4	25.3 / 52.9
rw_en	- / -	27.6 / 52.4	30.9 / 55.2	32.3 / 55.8	1.9 / 18.2	15.8 / 37.9	17.2 / 42.3	16.3 / 42.3
sd_en	- / -	30.6 / 54.2	34 / 57	34.1 / 56.9	1.5 / 12.1	15.8 / 40.8	14.6 / 40.8	13.3 / 42.1
shi_en	- / -	3.3 / 23.3	4.4 / 25.8	5.8 / 27.6	0.1 / 3.1	1.5 / 15.9	2.9 / 24	2.5 / 23.5
si_en	- / -	31.7 / 58	34 / 59.4	34.3 / 60	1.1 / 10.1	16.2 / 44.1	18.1 / 45.5	17.3 / 43.8
sk_en	- / -	40.7 / 65.6	42.5 / 66.6	42.5 / 66.5	8.7 / 31.5	26.6 / 52.2	33.1 / 58.4	35.1 / 61
sl_en	- / -	36.5 / 62.4	38.2 / 63.4	38.3 / 63.3	6.5 / 28.7	29.7 / 56.3	32.2 / 58.2	30.8 / 56.6
sm_en	- / -	25.6 / 50.5	28.9 / 53.6	28.5 / 53.5	0.6 / 9	6 / 24.1	14.2 / 39.9	14 / 37.7
sn_en	- / -	28.6 / 50.6	33.5 / 54.4	34.5 / 55	0.8 / 10.9	8.6 / 25.6	16.6 / 38.8	15.1 / 37.7
so_en	- / -	36.3 / 59.1	40.7 / 62.8	41.1 / 63.3	1.9 / 13.1	20.5 / 42.3	20.3 / 43.8	21.3 / 49
sq_en	- / -	41.9 / 65.8	43.5 / 66.8	43.5 / 66.9	12.6 / 35.4	35.4 / 60.4	35.6 / 60	36.9 / 62
sr_Latn_en	- / -	37 / 62.3	39.4 / 63.9	39 / 63.8	4.1 / 22.8	26.7 / 53.1	31.1 / 56.9	30.7 / 56
sr_en	- / -	37.3 / 62	39.2 / 63	39.6 / 63.5	5.4 / 24	27.7 / 53.6	31.3 / 57	33 / 58.6
ss_en	- / -	23.9 / 46.4	28.4 / 50.6	29.6 / 51.7	0.4 / 10	6.3 / 22.8	8.5 / 31.5	6.7 / 31.4
sv_en	- / -	45.5 / 68.1	46.9 / 69	46.8 / 68.9	8.2 / 31.9	40.7 / 64.3	41.7 / 64.9	42 / 65
sw_en	- / -	41.8 / 62.7	44.1 / 64.5	44.2 / 64.5	3.8 / 17.9	31 / 54.3	32.2 / 55.1	33.3 / 56.6
ta_en	- / -	27.3 / 53.5	28.4 / 54	29.4 / 54.9	1.3 / 7.4	15.9 / 42	15.9 / 43.5	18.2 / 45.6
te_en	- / -	27.5 / 53.7	28.6 / 54.4	29.6 / 55.1	1 / 4.8	12.5 / 36.6	18 / 44.4	19.2 / 46.4
tg_en	- / -	9.6 / 32.2	19.9 / 43.8	16.6 / 39	2.3 / 17.1	16.7 / 42	20.4 / 46.8	20.6 / 47.5
th_en	- / -	37.2 / 60.3	39.7 / 61.9	39.9 / 62.4	3.9 / 12	20 / 44.3	27.8 / 54.5	30.2 / 56.3
ti_en	- / -	17.5 / 42.2	22.2 / 46.7	24.1 / 48.4	0.7 / 10.2	8 / 29.9	8 / 31.9	9.4 / 31.8
tk_en	- / -	28.5 / 54.8	31.6 / 57	31 / 56.5	2.1 / 18.3	4.4 / 26.3	15.7 / 40.3	16 / 44
tn_en	- / -	24.2 / 48	29 / 52.8	30.9 / 53.8	0.3 / 9.2	7.6 / 29.1	9.6 / 32	10.8 / 34.9
to_en	- / -	15.9 / 39.9	21.2 / 45	21.4 / 44.7	0.3 / 6.8	7.2 / 25.8	7.6 / 26.3	3.2 / 26.5
tr_en	- / -	32.5 / 58.7	34.3 / 59.9	34.6 / 60	4.3 / 23.6	14.7 / 36.6	21.3 / 48.3	22.4 / 49.4
tt_en	- / -	14.4 / 36	14.4 / 34.9	19.3 / 41.7	1.1 / 9.7	9.5 / 38	16.8 / 42.6	16.9 / 43.7
ty_en	- / -	5.5 / 28.9	5.5 / 28.6	6.3 / 29.5	0.5 / 10.3	6 / 27.7	5.3 / 27	3.5 / 26.3
ug_en	- / -	9.1 / 32.2	14.7 / 39.9	13.4 / 35.6	1.2 / 13.8	10.5 / 35.6	12.6 / 39.4	11.6 / 37.7
uk_en	- / -	34 / 59.4	35.2 / 60.1	35.1 / 60.1	6.6 / 26.8	26.7 / 54.1	27.4 / 53.6	28.1 / 54.1
ur_en	- / -	3.4 / 18.9	3.8 / 19	3.7 / 19.1	4.3 / 20	22.2 / 49.7	23.9 / 50.5	25.6 / 53
uz_en	- / -	24.5 / 52.2	27.2 / 54.5	26.9 / 54.4	2.8 / 17.8	16.4 / 44.4	16.5 / 42.7	16.9 / 42.8
ve_en	- / -	19.1 / 41.9	21.2 / 44.6	24.3 / 46.7	0.2 / 9.3	5.4 / 21.9	4.7 / 25	3.3 / 22.2
vi_en	- / -	0.4 / 16	0.3 / 15.9	0.3 / 15.9	5.5 / 24.9	13.8 / 34.5	27.1 / 52.8	30.6 / 55.4
wo_en	- / -	2.1 / 18.4	2.4 / 19.3	2.8 / 19.4	0.1 / 7.3	2.2 / 20.9	2.2 / 20.3	2.2 / 20.2
xh_en	- / -	29.9 / 52.5	34.2 / 56.1	34.1 / 56.2	1 / 14.8	13.7 / 33.4	16.2 / 37.8	14.5 / 39.6
yo_en	- / -	13.3 / 37	14.5 / 37.4	20.9 / 43.5	0.4 / 11.1	3.8 / 23.9	3.4 / 21.5	4.6 / 23
yue_en	- / -	26.7 / 54.4	29.1 / 56.2	29.3 / 56.2	0.3 / 1.8	16.3 / 44.1	18.1 / 46	17.1 / 45.9
zh_Hant_en	- / -	27.1 / 54	29.1 / 55.5	29.1 / 55.6	1.6 / 7.7	10 / 29.8	19.2 / 46.4	19.5 / 46.6
zh_en	- / -	24.3 / 53.4	26.8 / 54.9	26.3 / 54.6	0.4 / 3	17.8 / 45.4	18.8 / 46.4	19.1 / 46.6
zu_en	- / -	6.5 / 24.3	7.5 / 25	7.5 / 25.1	0.9 / 13.2	13.1 / 32.6	17 / 38.6	16.7 / 40.4

xx2en	- / -	30.6 / 54.5	32.7 / 56.2	**33.6 / 57.6**	3.2 / 17.3	20.4 / 43.8	23.8 / 48.2	24.4 / 49.0
en2xx	- / -	16.5 / 39.6	17.6 / **41.9**	**17.9 / 41.9**	0.8 / 7.3	11.7 / 31.2	12.6 / 32.4	12.3 / 32.3
Average	19.8 / -	23.5 / 47.0	25.1 / 49.0	**25.7 / 49.7**	2.0 / 12.3	16.0 / 37.4	18.1 / 40.2	18.3 / 40.6

A.11 Datasheet

Datasheet - `MADLAD-400`

Motivation

1. For what purpose was the dataset created? (Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.) *We create* `MADLAD-400` *as a general purpose monolingual document level dataset covering 419 languages for the purpose of providing general training data for multilingual NLP tasks such as MT and language modeling. One of the goals of the Language Inclusivity Moonshot (LIM)* [2] *is to scale our language support to 1,000 languages and to support speech recognition for 97% of the global population. A core objective associated with this goal is to open source data and models. Our expectation is that releasng* `MADLAD-400` *will foster progress on the language research, especially on medium and low resource languages. An estimate of over 1B people globally speak languages that are not covered by mainstream models at Google or externally.*

2. Who created this dataset and on behalf of which entity? *Sneha Kudugunta*[†]*, Isaac Caswell*[◇]*, Biao Zhang*[†]*, Xavier Garcia*[†]*, Derrick Xin*[†]*, Aditya Kusupati*[◇]*, Romi Stella*[†]*, Ankur Bapna*[†]*, Orhan Firat*[†] *(*[†]*Google DeepMind,* [◇]*Google Research*)

3. Who funded the creation of the dataset? *Google Research and Google DeepMind*

4. Any other comments? *None*

Composition

1. What do the instances that comprise the dataset represent? *Each instance is a preprocessed web-crawled document whose language that we annotated using a LangID model described by (Caswell et al., 2020). For the sentence level version, we used a sentence-splitter to split the documents into sentences and then deduplicated the resulting dataset.*

2. How many instances are there in total? `MADLAD-400` *has 4.0B documents (100B sentences, or 2.8T tokens) total across 419 languages with the median language containing 1.7k documents (73k sentences of 1.2M tokens.)*

---

[2] `https://blog.google/technology/ai/ways-ai-is-scaling-helpful/`

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? (If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).) *MADLAD-400 are created from CommonCrawl documents that have been annotated by language, filtered and preprocessed. To maintain high precision, we filtered out data aggressively, and may not have captured every document of a given language in CommonCrawl. Moreover, there may also be languages in CommonCrawl that we may not have mined.*

4. What data does each instance consist of? *Each instance is raw text in either document form for the document level data, or in sentence form for the sentence level data.*

5. Is there a label or target associated with each instance? If so, please provide a description. *No.*

6. Is any information missing from individual instances? *No.*

7. Are relationships between individual instances made explicit? *No.*

8. Are there recommended data splits (e.g., training, development/validation, testing)? *No.*

9. Are there any errors, sources of noise, or redundancies in the dataset? *While we have taken extensive care to audit and filter MADLAD-400, there may still be documents annotated with the wrong language or documents of low quality.*

10. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? (If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.) *Yes*

11. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public

communications)? (If so, please provide a description.) *Given that MADLAD-400 is a general web-crawled dataset it is possible that documents in the dataset may contain such information.*

12. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? (If so, please describe why.) *Given that MADLAD-400 is a general web-crawled dataset, even after filtering, it is possible that there are documents containing offensive content, etc.*

13. Does the dataset relate to people? *It is likely that some documents in MADLAD-400 contain sentences referring to and describing people.*

14. Does the dataset identify any subpopulations (e.g., by age, gender)? *It is likely that some documents in MADLAD-400 contain sentences referring to and describing people of certain subpopulations.*

15. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? *Yes, it is possible that their names are mentioned in certain documents.*

16. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? *Given that MADLAD-400 is a general web-crawled dataset, even after filtering, it is possible that there are documents containing sensitive data.*

17. Any other comments? *None.*

Collection

1. How was the data associated with each instance acquired? *Each instance was acquired by performing transformations on the documents in all available snapshots of CommonCrawl as of August 20, 2022.*

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? *We annotated the CommonCrawl data using a LangID model trained using the procedure described by (Caswell et al., 2020). Then, we manually inspected the data and then filtered or preprocessed the documents to create MADLAD-400.*

3. If the dataset is a sample from a larger set, what was the sampling strategy? *MADLAD-400 is a subset of CommonCrawl documents determined using LangID annotations and filtering/preprocessing steps.*

4. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? *For the audit, the authors inspected the dataset. In some cases native speaker volunteers provided advice on the quality of the dataset.*

5. Over what timeframe was the data collected? (Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.) *We do not annotate timestamps. The version of CommonCrawl that we used has webcrawls ranging from 2008 to August 2022.*

6. Were any ethical review processes conducted (e.g., by an institutional review board)? *No.*

7. Does the dataset relate to people? *It is likely that some documents in MADLAD-400 contain sentences referring to and describing people.*

8. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? *We collected this data via webpages crawled by CommonCrawl.*

9. Were the individuals in question notified about the data collection? *No.*

10. Did the individuals in question consent to the collection and use of their data? *No.*

11. If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? *No.*

12. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? *No.*

13. Any other comments? *None.*

Preprocessing/cleaning/labeling

1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? (If so,

please provide a description. If not, you may skip the remainder of the questions in this section.) *Various types of preprocessing were done: deduplication of 3 sentence spans, filtering substrings according to various heuristics associated with low quality, Virama encoding correction, converting Zawgyi encoding to Unicode encoding for Myanmar script characters and a Chinese pornographic content filter heuristic. In addition, 79 annotated language datasets were removed on inspection due to low quality or mislabeling.*

2. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? *No, this raw data is hosted by CommonCrawl.*

3. Is the software used to preprocess/clean/label the instances available? *As of June 13, 2023, no.*

4. Any other comments? *None.*

Uses

1. Has the dataset been used for any tasks already? (If so, please provide a description.) *MADLAD-400 has been used for MT and language modeling.*

2. Is there a repository that links to any or all papers or systems that use the dataset? *No*

3. What (other) tasks could the dataset be used for? *This dataset could be used as a general training dataset for any of the languages in MADLAD-400.*

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? (For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?) *While steps have been taken to clean MADLAD-400, content containing sensitive content about individuals or groups could affect the performance of some downstream NLP tasks. Moreover, while building applications for (a) given language(s), we urge practitioners to assess the suitability of MADLAD-400 for their usecase.*

5. Are there tasks for which the dataset should not be used? (If so, please provide a description.) *N/A.*

6. Any other comments? *None.*

Distribution

1. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? *MADLAD-400 is made available through a GCP bucket.*
2. When will the dataset be distributed? *June 2023*
3. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? (If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.) *AI2 has made a version of this data available under the ODC-BY license. Users are also bound by the CommonCrawl terms of use in respect of the content contained in the dataset.*
4. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? *Users are bound by the CommonCrawl terms of use in respect of the content contained in the dataset.*
5. Any other comments? *None.*

Maintenance

1. Who is supporting/hosting/maintaining the dataset? *An external organization, AI2 is hosting the dataset.*
2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)? *Sneha Kudugunta snehakudugunta@google.com or Isaac Caswell (icaswell@google.com) for questions about the dataset contents, or Dirk Groeneveld dirkg@allenai.org for questions related to the hosting of the dataset.*
3. Is there an erratum? (If so, please provide a link or other access point.) *No*
4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances')? (If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?) *There are no such plans, but major issues may be corrected when reported through email or the Github page.*
5. If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? (If so, please describe these limits and explain how they will be enforced.) *N/A*

6. Will older versions of the dataset continue to be supported/hosted/maintained? (If so, please describe how. If not, please describe how its obsolescence will be communicated to users.) *No*

7. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? (If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.) *A relatively unprocessed version of* `MADLAD-400`*,* `MADLAD-400-noisy` *is made available for others to build upon using superior cleaning/preprocessing techniques for their specific usecases.*

8. Any other comments? *None*

A.12 Model Card

Model Card

Model Details

- Person or organization developing model: *Google DeepMind and Google Research*
- Model Date: *June 13, 2023*
- Model Types: *Machine Translation and Language Modeling Models.*
- Information about training algorithms, parameters, fairness constraints or other applied approaches, and features: *Provided in the paper.*
- Paper: *Kudugunta et al,* `MADLAD-400`*: Monolingual And Document-Level Large Audited Dataset, Under Review, 2023*
- License: *ODC-BY*
- Contact: *snehakudugunta@google.com*

Intended Use

- Primary intended uses: *Machine Translation and multilingual NLP tasks on over 400 languages.*

- Primary intended users: *Research community.*
- Out-of-scope use cases: *These models are trained on general domain data and are therefore not meant to work on domain-specific models out-of-the box. Moreover, these research models have not been assessed for production usecases.*

Factors

- *The translation quality of this model varies based on language, as seen in the paper, and likely varies on domain, though we have not assessed this.*

Metrics

- *We use SacreBLEU and chrF, two widely used machine translation evaluation metrics for our evaluations.*

Ethical Considerations

- *We trained these models with* `MADLAD-400` *and publicly available data to create baseline models that support NLP for over 400 languages, with a focus on languages underrepresented in large-scale corpora. Given that these models were trained with web-crawled datasets that may contain sensitive, offensive or otherwise low-quality content despite extensive preprocessing, it is still possible that these issues to the underlying training data may cause differences in model performance and toxic (or otherwise problematic) output for certain domains. Moreover, large models are dual use technologies that have specific risks associated with their use and development. We point the reader to surveys such as those written by Weidinger et al. (2021) or Bommasani et al. (2021) for a more detailed discussion of these risks, and to Liebling et al. (2022) for a thorough discussion of the risks of machine translation systems.*

Training Data

- *For both the machine translation and language model,* `MADLAD-400` *is used. For the machine translation model, a combination of parallel datasources covering 157 languages is also used. Further details are described in the paper.*

Evaluation Data

- *For evaluation, we used WMT, NTREX, Flores-200 and Gatones datasets as described in Section 2.1.4.3.*

Caveats and Recommendations

- *We note that we evaluate on only 204 of the languages supported by these models and on machine translation and few-shot machine translation tasks. Users must consider use of this model carefully for their own usecase.*

Table A.7: Evaluation scores on WMT (depicted as <bleu> / <chrf>) for the MT models and language models described in Section 2.1.4.1 and Section 2.1.4.2 compared against NLLB-54B.

	NLLB	**MT-3B**	**MT-7.2B**	**MT-10.7B**	**LM-8B**			
					0-shot	1-shot	5-shot	10-shot
csen	33.6 / 58.8	34.8 / 59.9	35.7 / 60.5	35.3 / 60.4	3 / 22.2	27.6 / 53.7	26.7 / 51.8	28.2 / 54
deen	37.7 / 62.4	37.3 / 62.5	38.1 / 63	38 / 63	3.3 / 23.6	26 / 51.9	29.7 / 55.2	30.1 / 55.5
esen	37.6 / 61.7	38.1 / 62.4	38.5 / 62.6	38.5 / 62.7	2.2 / 19	30 / 55.2	31.5 / 57.4	32.3 / 58.4
eten	34.7 / 61.1	36.2 / 61.9	37.4 / 62.8	37.4 / 62.9	4.7 / 24.7	26.4 / 52.8	27.9 / 55	27.7 / 54.5
fien	28.8 / 55.7	32.8 / 59.1	33.5 / 59.4	33.3 / 59.6	1.6 / 16.8	27.1 / 53.5	26.2 / 52	26.3 / 52.6
fren	41.9 / 65.7	42.2 / 65.6	42.7 / 65.9	42.4 / 65.8	2 / 18.6	28.9 / 54.8	35.1 / 60.7	35.6 / 61.1
guen	31.2 / 58.9	30.4 / 58.2	29.9 / 58	29.9 / 57.9	1 / 13	22.2 / 50.7	21.9 / 51.2	24.1 / 53.2
hien	37.4 / 64.2	33.5 / 61.3	36.1 / 62.7	35.5 / 62.4	1.2 / 10.2	23.4 / 52.3	23.4 / 54.5	24.4 / 54.5
kken	30.2 / 58.4	12.3 / 48.1	22.9 / 49.8	19.9 / 52.1	2.4 / 14.2	15.3 / 41.2	11.5 / 40.9	11.9 / 41.6
lten	29.7 / 58.8	34.5 / 60.9	35.3 / 61.4	35.1 / 61.5	3 / 21.4	24.1 / 51.1	26.1 / 52.8	28.1 / 54.9
lven	24.8 / 53.1	27.7 / 55.6	28.1 / 56	28.3 / 56.1	1.3 / 15.7	21.4 / 49.3	21.4 / 49.2	22 / 50.1
roen	43.4 / 66.4	43.6 / 66.6	44.4 / 67	44.4 / 67.1	4.1 / 24.8	34.9 / 58.9	36.1 / 59.9	36.2 / 60.5
ruen	39.9 / 63.8	39.9 / 64.4	40.7 / 64.9	40.5 / 64.9	2.1 / 16.4	32.7 / 56.9	34.1 / 58.5	27.4 / 53.7
tren	34.3 / 60.5	33.2 / 58.9	34.3 / 59.7	34.1 / 59.8	1.6 / 16.8	21.7 / 48.1	21.7 / 47.7	22.1 / 48.7
zhen	28.5 / 56.4	24.9 / 55.6	26.2 / 56	25.7 / 56.2	0.4 / 1.6	15.3 / 40	19.3 / 46.2	16.9 / 47
encs	25.2 / 53.2	26.2 / 53.8	27 / 54.2	26.8 / 54.2	0.7 / 10.3	18.5 / 45.2	19.3 / 48	19.6 / 47.9
ende	33 / 62	33.7 / 62.5	35 / 63.2	35 / 63.2	1.6 / 16.2	19.7 / 48.1	21.9 / 53	21.9 / 54.3
enes	37.2 / 61.1	37.3 / 61.1	37.7 / 61.3	37.4 / 61.3	0.9 / 13.9	28.4 / 53.7	30.2 / 55.6	30.9 / 56.3
enet	27 / 59.3	27.6 / 60.4	28.3 / 61	28.4 / 61.1	0.5 / 10.7	20.1 / 52.5	19.9 / 51.2	20.8 / 53.4
enfi	27.7 / 61.7	28.3 / 60	29 / 60.4	27.8 / 60.6	0.5 / 10.5	18.1 / 49	19.8 / 51.6	20.2 / 51.9
enfr	44.2 / 67.7	44.8 / 67.2	45.3 / 67.6	45.6 / 67.7	1 / 13.7	32.8 / 58.6	32.1 / 57.6	32.7 / 59.5
engu	17.6 / 50	18.6 / 51.3	19.2 / 52.1	18.8 / 52	0.1 / 0.3	0.2 / 1.2	0.1 / 1.4	0.1 / 1.6
enhi	26 / 53.5	23.4 / 50.1	24.2 / 50.8	24.5 / 51.1	1.5 / 10.6	16.9 / 42.6	15.7 / 40.5	16.7 / 42.7
enkk	34.8 / 65.2	7.6 / 44.2	12.9 / 47	10.7 / 47.1	0.2 / 5.2	6.1 / 37.6	6.2 / 37.7	5.7 / 38.9
enlt	37 / 66.9	26.5 / 59	27.2 / 59.5	27 / 59.6	1.1 / 14.2	20.6 / 53.3	20.9 / 53.6	20.1 / 53.1
enlv	21.3 / 53.6	23.9 / 55.8	25.2 / 56.6	24.8 / 56.4	1 / 12.4	18.3 / 48.5	19.7 / 52	19.8 / 51.6
enro	33.4 / 60.4	35.5 / 61.4	35.5 / 61.5	35.6 / 61.7	3.7 / 21	29.6 / 56.5	25.3 / 54	29.5 / 56.4
enru	44.8 / 67.4	32.9 / 57.8	33.6 / 58.4	33.6 / 58.3	0.6 / 2.5	20.4 / 44	22.3 / 47	23.8 / 49.4
entr	23.3 / 58.3	24.2 / 57.2	26.1 / 58.3	25.9 / 58.1	0.5 / 10.9	11.1 / 42.2	11 / 46.4	8.6 / 46.1
enzh	33.9 / 30.1	32.6 / 29.8	33.3 / 30.9	33.6 / 31.1	0.9 / 2.7	20.2 / 18.9	16.9 / 18.1	19 / 19
xx2en	34.2 / 60.4	33.4 / 60.0	34.9 / 60.6	**34.6 / 60.8**	2.3 / 17.3	25.1 / 51.4	26.2 / 52.9	26.2 / 53.4
en2xx	**31.1 / 58.0**	28.2 / 55.4	29.3 / 56.2	29.0 / 56.2	1.0 / 10.3	18.7 / 43.5	18.8 / 44.5	19.3 / 45.5
Average	**32.7 / 59.2**	30.8 / 57.7	32.1 / 58.4	31.8 / 58.5	1.6 / 13.8	21.9 / 47.4	22.5 / 48.7	22.8 / 49.4

Appendix B

(MIS)FITTING: SUPPLEMENTARY MATERIALS

B.1 Our model training (§3.1.9)

We train a variety of Transformer LMs, ranging in size from 12 million to 1 billion parameters, on FineWeb (Penedo et al., 2024), tokenized with the GPT-NeoX-20B tokenizer (Black et al., 2022), which is a BPE tokenizer with a vocabulary size of 50257, trained on the Pile (Gao et al., 2020). All models were trained on a combination of NVIDIA GeForce RTX 2080 Ti, Quadro RTX 6000, NVIDIA A40, and NVIDIA L40 GPUs. These transformers follow the standard architecture, and use pre-layer RMSNorm (Kudo, 2018), the SwiGLU activation function (Shazeer, 2020), and rotary positional embeddings (Su et al., 2024). We use a batch size of 512 with a sequence length of 2048. The learning rate is warmed up linearly over 50 steps to the peak learning rate and then follows a cosine decay to 10% of the peak. We use an Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.95$. We sweep over learning rates and data budgets at each model scale. For our evaluation metric, we use perplexity on a validation set of C4 (Raffel et al., 2020). See Table B.1 and Table B.8 for additional architecture details and hyperparameters, including data budget and learning rate.

Hyperparameter	Value
Vocabulary size	50K
Batch Size	512
Sequence Length	2048
Attention Head Size	64
Learning Rate	Swept $2^{\{0,1,2,3\}} \cdot 10^{\{-3,-4\}}$
Feedforward Dimension	4 x hidden dimension
LR Schedule	Cosine Decay
LR Warmup	50 steps
End LR	0.1 x Peak LR
Optimizer	Adam $(\beta_1, \beta_2 = 0.9, 0.95)$

Table B.1: Hyperparameter details for models trained in §3.1.9

Model size	# layers	Hidden Dim	# Attn Heads	# Steps
12M	5	448	7	{100, 200, 250, 360, 500, 750, 1,000, 4,000}
17M	7	448	7	{100, 200, 250, 500, 750, 1,000, 1,250, 1,500}
25M	8	512	8	{250, 360, 500, 750, 1,000, 1,500, 2,000, 8,000, 16,000}
35M	9	576	9	{200, 250, 360, 500, 750, 1,000, 1,250, 1,500, 4,000, 16,000}
50M	10	640	10	{250, 500, 750, 1,000, 1,250, 1,500, 1,800, 2,000, 2,500, 4,000, 16,000}
70M	12	704	11	{250, 500, 750, 1,000, 1,250, 1,500, 2,000, 2,500, 3,000, 5,000}
100M	14	768	12	{250, 500, 1,000, 1,500, 2,000, 2,500, 3,000, 4,000, 6,000, 12,000}
200M	18	960	15	{500, 1,500, 6,000, 9,000}
300M	19	1152	18	{4,000}
400M	20	1280	20	{3,000, 6,000}
1B	26	1792	28	{20,000}

Table B.2: Architecture and Data budget details for models trained in §3.1.9

B.2 Full Checklist

We define each category as follows:

- **Scaling Law Hypothesis:** This specifies the form of the scaling law, that of the variables and parameters, and the relation between each.
- **Training Setup:** This specifies the exact training setup of each of the models trained to test the scaling law hypothesis.
- **Data Collection:** Evaluating various checkpoints of our trained models to collect data points that will be used to fit a scaling law in the next stage.
- **Fitting Algorithm:** Using the data points collected in the previous stage to optimize the scaling law hypothesis.

Scaling Law Reproducilibility Checklist

Scaling Law Hypothesis (§3.1.3)

- What is the form of the power law?
- What are the variables related by (included in) the power law?
- What are the parameters to fit?
- On what principles is this form derived?
- Does this form make assumptions about how the variables are related?

Training Setup (§3.1.6)

- How many models are trained?
- At which sizes?
- On how much data each? On what data? Is any data repeated within the training for a model?

- How are model size, dataset size, and compute budget size counted? For example, how are parameters of the model counted? Are any parameters excluded (e.g., embedding layers)? How is compute cost calculated?
- Are code/code snippets provided for calculating these variables if applicable?
- How are hyperparameters chosen (e.g., optimizer, learning rate schedule, batch size)? Do they change with scale?
- What other settings must be decided (e.g., model width vs. depth)? Do they change with scale?
- Is the training code open source?

Data Collection(§3.1.7)

- Are the model checkpoints provided openly?
- How many checkpoints per model are evaluated to fit each scaling law? Which ones, if so?
- What evaluation metric is used? On what dataset?
- Are the raw evaluation metrics modified? Some examples include loss interpolation, centering around a mean or scaling logarithmically.
- If the above is done, is code for modifying the metric provided?

Fitting Algorithm (§3.1.8)

- What objective (loss) is used?
- What algorithm is used to fit the equation?
- What hyperparameters are used for this algorithm?
- How is this algorithm initialized?
- Are all datapoints collected used to fit the equations? For example, are any outliers dropped? Are portions of the datapoints used to fit different equations?

- How is the correctness of the scaling law considered? Extrapolation, Confidence Intervals, Goodness of Fit?

B.3 Full Sheet

We provide an overview of all the papers surveyed in Tables B.3,B.4, B.5, B.6 and B.7.

Table B.3: Details on domain of experiments and availability of code by category for each paper surveyed.

Paper	Domain	Training Code?	Analysis Code?	Checkpoints?	Metric Scores?
Rosenfeld et al. (2019)	Vision, LM	N	N	N	N
Mikami et al.	Vision	N	Y	Y	Y
Schaeffer et al. (2023)	LM	N	N	N	N
Sardana & Frankle (2023)	LM	N	N	N	N
Sorscher et al. (2022)	Vision	N	N	N	Y
Caballero et al. (2022)	LM	N	Y	N	Y
Besiroglu et al. (2024)	LM		Y	N	Y
Gordon et al. (2021)	NMT	Y	Y	Y	Y
Bansal et al. (2022)	NMT	N	N	N	N
Hestness et al. (2017)	NMT, LM, Vision, Speech	N	N	N	N
Bi et al. (2024)	LM	N	N	N	N
Bahri et al. (2024)	Vision	N	N	N	N
Geiping et al. (2022)	Vision	Y	Y	N	N
Poli et al. (2024)	LM	N	N	N	N
Hu et al. (2024)	LM	Y	N	N	N
Hashimoto (2021)	NLP	N	N	N	N
Ruan et al. (2024)	LM	Y	Y	N	Y
Anil et al. (2023)	LM	N	N	N	N
Pearce & Song (2024a)	LM	N	Y	N	N
Cherti et al. (2023)	VLM	Y	Y	Y	Y
Porian et al. (2024)	LM	Y	Y	N	Y
Alabdulmohsin et al. (2022)	LM, Vision	N	Y	Y	Y
Gao et al. (2024)	NLP	Y	Y	Y	N
Muennighoff et al. (2024)	LM	Y	Y	Y	N
Rae et al. (2021)	LM	N	N	N	N
Shin et al. (2023)	RecSys	N	N	N	N
Hernandez et al. (2022)	LM	N	N	N	N
Filipovich et al. (2022)	LM	N	N	N	N
Neumann & Gros (2022)	RL	Y	Y*	Y	N
Droppo & Elibol (2021)	Speech	N	N	N	N
Henighan et al. (2020)	LM, Vision, Video, VLM	N	N	N	N
Goyal et al. (2024)	LM, Vision, VLM	N	Y	N	Y
Aghajanyan et al. (2023)	Multimodal LM	N	N	N	N
Kaplan et al. (2020)	LM	N	N	N	N
Ghorbani et al. (2021)	NMT	N	Y	N	N
Gao et al. (2023)	RL/LM	N	N	N	N
Hilton et al. (2023)	RL	N	N	N	N

Continued on next page

Table B.3 continued from previous page

Paper	Domain	Training Code?	Analysis Code?	Checkpoints?	Metric Scores?
Frantar et al. (2023)	LM, Vision	N	N	N	N
Prato et al. (2021)	Vision	Y*	Y	Y	Y
Covert et al. (2024)	LM	Y	Y	N	N
Hernandez et al. (2021)	LM	N	N	N	N
Ivgi et al. (2022)	NLP	N	N	N	N
Tay et al. (2022a)	LM	N	N	N	N
Tao et al. (2024)	LM	N	Y	N	Y
Jones (2021)	RL	Y	Y	N	Y
Zhai et al. (2022)	Vision	Y	N	N	N
Dettmers & Zettlemoyer (2023)	LM	N	N	N	N
Dubey et al. (2024)	LM	N	N	N	N
Hoffmann et al. (2022)	LM	N	N	N	N
Ardalani et al. (2022)	RecSys	N	N	N	N
Clark et al. (2022)	LM	N	Y	N	Y

Paper	Power Law Form	Purpose Of Power Law (E.G., Performance Prediction, Optimal Ratio)	# Power Law Parameters	# Of Scaling Laws
Rosenfeld et al. (2019)	$\tilde{e}(m,n) = an^{-\alpha} + bm^{-\beta} + c_\infty$	Performance Prediction	5-6	8
Mikami et al.	$L(n,s) = \delta\left(\gamma + n^{-\alpha}\right)s^{-\beta}$	Performance Prediction	4	3
Schaeffer et al. (2023)	None	N/A	NA	NA
Sardana & Frankle (2023)	$L(N,D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}$; $N^*\left(\ell, D_{\text{inf}}\right), D_{\text{tr}}^*\left(\ell, D_{\text{inf}}\right) = \arg\min_{N,D_{\text{tr}}\mid L(N,D_{\text{tr}})=\ell}$	Performance Prediction	5	4
Sorscher et al. (2022)	$c\cdot\alpha^{-\beta}, c\cdot\exp(-b\alpha)$	Performance Prediction	2	34
Caballero et al. (2022)	$y = a + (bx^{-c_0})\prod_{i=1}^{n}\left(1+\left(\frac{x}{d_i}\right)^{1/f_i}\right)^{-c_i * f_i}$	Performance Prediction	5+	100+
Besiroglu et al. (2024)	$L(N,D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}$	Performance Prediction	5	1
Gordon et al. (2021)	$L(N,D) = \left[\left(\frac{N}{N_c}\right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D}{D_c}\right]^{\alpha_D}$	Performance Prediction	4	3
Bansal et al. (2022)	$L(D) = \alpha\left(D^{-1}+C\right)^p$	Performance Prediction	3	20
Hestness et al. (2017)	$\varepsilon(m) \sim \alpha m^{\beta_g} + \gamma$	Performance Prediction	3	17
Bi et al. (2024)	$M_{\text{opt}} = M_{\text{base}}\cdot C^a\ \ \eta_{\text{opt}} = 0.3118\cdot C^{-0.1250}$ $D_{\text{opt}} = D_{\text{base}}\cdot C^b\ \ B_{\text{opt}} = 0.2920\cdot C^{0.3271}$	Optimal Ratio, Performance Prediction	2	5
Bahri et al. (2024)	$L(D)\propto D^{-\alpha_K},\quad L(P)\propto P^{-\alpha_K}$	Performance Prediction	2	35
Geiping et al. (2022)	$f(x) = ax^{-c}+b$, $v_{\text{Effective Extra Samples from Augmentations}}(x) = f_{\text{ref}}^{-1}\left(f_{\text{aug}}(x)\right) - x$	Performance Prediction	3	50
Poli et al. (2024)	$\log N^* \propto a\log C$ and $\log D^* \propto b\log C$	Performance Prediction	2	
Hu et al. (2024)	$L(N,D) = C_N N^{-\alpha} + C_D D^{-\beta} + L_0$	Performance Prediction	5	6
Hashimoto (2021)	$\min_{\lambda,\alpha}\mathbb{E}_{\hat{q},\hat{n}}\left[\left(\log(R(\hat{n},\hat{q})-\epsilon) - \alpha\log(\hat{n}) + \log\left(C_\lambda(\hat{q})\right)\right)^2\right]$ $R(\hat{n},\hat{q}) = \mathbb{E}\left[\ell\left(\hat{\theta}\left(p_{\hat{n},\hat{q}}\right);x,y\right)\right]$	Performance Prediction	2+n(data mixes)	4
Ruan et al. (2024)	$E_m \approx h\sigma\left(\beta^\top S_m + \alpha\right)$	Performance Prediction	3	
Anil et al. (2023)	$N^*(C)\approx N_0^*\cdot C^a$	Performance Prediction	2	1
Pearce & Song (2024a)	$N_{\backslash E} = bC_{\backslash E}^m\ L = bC^m$	Optimal Ratio, Performance Prediction	2	1
Cherti et al. (2023)	$E = \beta C^\alpha$	Performance Prediction	2	8
Porian et al. (2024)	$N^*(C)\approx N_0^*\cdot C^a$	Optimal Ratio	2	6
Alabdulmohsin et al. (2022)	$\varepsilon_x = \beta x^c$; $\varepsilon_x - \varepsilon_\infty = \beta x^c$; $\varepsilon_x = \beta\left(x^{-1}+\gamma\right)^{-c}$; $\varepsilon_x = \gamma(x)(1+\gamma(x))^{-1}\varepsilon_0 + (1+\gamma(x))^{-1}\varepsilon_\infty$	Performance Prediction	2-4	600
Gao et al. (2024)	$L(n,k) = \exp\left(\alpha+\beta_k\log(k)+\beta_n\log(n)+\gamma\log(k)\log(n)\right) + \exp(\zeta+\eta\log(k))$	Performance Prediction	2-6	1
Muennighoff et al. (2024)	$L\left(U_N,U_D,R_N,R_D\right) = \frac{A}{\left(U_N+U_N R_N^*\left(1-e^{\frac{-R_N}{R_N^*}}\right)\right)^\alpha} + \frac{B}{\left(U_D+U_D R_D^*\left(1-e^{\frac{-R_D}{R_D^*}}\right)\right)^\beta} + E$	Performance Prediction	2 (+4)	1
Rae et al. (2021)	None	Performance Prediction	N/A	N/A
Shin et al. (2023)	None	Scaling trend	NA	NA
Hernandez et al. (2022)	$E = k * N^\alpha$	Optimal Ratio	2	1
Filipovich et al. (2022)	$\mathcal{L}(C) = (C_e C)^{\alpha_C}$	Performance Prediction	2	3
Neumann & Gros (2022)	$N_{\text{opt}}(C) = \left(\frac{C}{C_0}\right)^{\alpha_C^{\text{opt}}}$, $E_i = \frac{1}{1+(N_j/N_i)^{\alpha_N}}$	Performance Prediction	2	3 * 2
Droppo & Elibol (2021)	$L(N,D) = \left[\left(L_\infty\right)^{\frac{1}{\alpha}} + \left(\frac{N_C}{N}\right)^{\frac{\alpha_N}{\alpha}} + \left(\frac{D_C}{D}\right)^{\frac{\alpha_D}{\alpha}}\right]^\alpha$	Performance Prediction	6	3
Henighan et al. (2020)	$L(x) = L_\infty + \left(\frac{x_0}{x}\right)^{\alpha_x}$	Performance Prediction	3	36
Goyal et al. (2024)	$y_k = a\cdot n_1^{b_1}\prod_{j=2}^{k}\left(\frac{n_j}{n_{j-1}}\right)^{b_j} + d$	Performance Prediction	2+ 2*n(data mixes)	1
Aghajanyan et al. (2023)	$L(N,D_j) = E_j + \frac{A_j}{N^{\alpha_j}} + \frac{B_j}{\lvert D_j\rvert^{\beta_j}}$, $L(N,D_i,D_j) = [\frac{L(N,D_i)+L(N,D_j)}{2}] - C_{i,j} + \frac{A_{i,j}}{N^{\alpha_{i,j}}} + \frac{B_{i,j}}{\lvert D_i\rvert+\lvert D_j\rvert^{\beta_{i,j}}}$	Performance Prediction	5	14
Kaplan et al. (2020)	$L(N,D) = \left[\left(\frac{N}{N_c}\right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D}{D_c}\right]^{\alpha_D}$	Performance Prediction	4	7
Ghorbani et al. (2021)	$\text{BLEU} = c_B L^{-p_B}$, $\hat{L}_{opt}(B) = \alpha^* B^{-(p_d+p_e)} + L_\infty,\quad \alpha^* \equiv \alpha\left(\frac{S_e(p_e+p_d)}{p_e}\right)^{p_e}\left(\frac{S_d(p_e+p_d)}{p_d}\right)^{p_d}$	Optimal Ratio, Performance Prediction	6	8
Gao et al. (2023)	$R_{\text{bon}}(d) = d\left(\alpha_{\text{bon}} - \beta_{\text{bon}}d\right)$, $R_{\text{RL}}(d) = d\left(\alpha_{\text{RL}} - \beta_{\text{RL}}\log d\right)$	Performance Prediction	2	2
Hilton et al. (2023)	$I^{-\beta} = \left(\frac{N}{N}\right)^{\alpha_N} + \left(\frac{E_i}{E}\right)^{\alpha_E}$	Optimal Ratio, Performance Prediction	5	3
Frantar et al. (2023)	$L(S,N,D) = \left(a_S(1-S)^{k_S} + c_S\right)\cdot\left(\frac{1}{N}\right)^{b_N} + \left(\frac{a_D}{D}\right)^{b_D} + c$	Optimal Ratio, Performance Prediction	7	2
Prato et al. (2021)	$\text{Err}(N) = \text{Err}_\infty + kN^\alpha$, $\text{Err}(C) = \text{Err}_\infty + kC^\alpha$,	Performance Prediction	3	12
Covert et al. (2024)	$\log\lvert\psi_k(z)\rvert \approx \log\lvert c(z)\rvert - \alpha(z)\log(k)$	Performance Prediction	2	Many
Hernandez et al. (2021)	$L \approx \left[\left(\frac{N_C}{N}\right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D_C}{k(D_F)^\alpha(N)^\beta}\right]^{\alpha_D}$	Performance Prediction	3	1
Ivgi et al. (2022)	NS	Performance Prediction	NA	NA
Tay et al. (2022a)	None	Scaling trend	NA	NA
Tao et al. (2024)	$N_{\text{v}}^{\text{opt}} = N_{\text{v}}^0 * \left(\frac{N_{\text{ve}}}{N_{\text{ve}}^0}\right)^\gamma$, $\mathcal{L}_{\text{u}} = -E + \frac{A_1}{N_{\text{ue}}^{\alpha_1}} + \frac{A_2}{N_{\text{v}}^{\alpha_2}} + \frac{B}{D^\beta}$	Optimal Ratio, Performance Prediction	7	2
Jones (2021)	$\text{plateau} = m_{\text{boardsize}}^{\text{plateau}}\cdot\text{boardsize} + c^{\text{plateau}}$ $\text{incline} = m_{\text{boardsize}}^{\text{incline}}\cdot\text{boardsize} + m_{\text{flops}}^{\text{incline}}\cdot\log\text{flop} + c^{\text{incline}}$ $\text{elo} = \text{incline.clamp}(\text{plateau}, 0)$	Performance Prediction	5	1
Zhai et al. (2022)	$E = \alpha + \beta(C+\gamma)^{-\mu}$	Performance Prediction	4	3
Dettmers & Zettlemoyer (2023)	None	Scaling trend	NA	NA
Dubey et al. (2024)	$N^*(C) = AC^\alpha$.	Optimal Ratio	2	2
Hoffmann et al. (2022)	A3: $L(N,D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}$	Optimal Ratio, Performance Prediction	5	3
Ardalani et al. (2022)	$\left(\alpha x^{-\beta}+\gamma\right)$	Performance Prediction	3	3
Clark et al. (2022)	$\log L(N,E) \triangleq +a\log N + b\log E + c\log N\log E + d$	Performance Prediction	4	3

Table B.4: Details on power law for each paper surveyed.

Paper	Training Runs / Law	Max. Training Flops	Max. Training Params	Max. Training Data	Data Described?	Hyperparameters Described?	How Are Model Params Counted (E.G., W/ Or W/Out Embeddings)
Rosenfeld et al. (2019)	42-49		0.7M-70M	100M words / 1.2M images	Y	Y	Non-embedding
Mikami et al.	7		ResNet-101	64k-1.28M images	Y	Y	NA
Schaeffer et al. (2023)	4		10^{11}	NA	Y	NA	Non-embedding
Sardana & Frankle (2023)	47		150M-6B	1.5B-1.25T tokens	N	Y	NA
Sorscher et al. (2022)	60		86M (ViT)	200 epochs	Y	Y	NA
Caballero et al. (2022)	3-40		NS	NS	N	N	NS
Besiroglu et al. (2024)	NA	NA	NA	NA	Y	NA	Non-embedding
Gordon et al. (2021)	45-55		56M	28.3M-51.1M examples	Y	Y	Non-embedding
Bansal et al. (2022)	10		170M-800M	500K-512M sentences (28B tokens)	Y	Y	NS
Hestness et al. (2017)	9		upto 193M	$2^{19} - 2^{28}$ tokens, upto 2^9 images, $2k$ audio hours	Y	Y	NS
Bi et al. (2024)	80	$1e17 - 3e20$			Y	Y	Non-embedding
Bahri et al. (2024)	8-27		36.5M	upto 78k steps; 100 epochs	Y	Y	NS
Geiping et al. (2022)	13		ResNet-18	upto 7.6M images	Y	Y	NS
Poli et al. (2024)	500 total	8.00E+19	70M-7B		Y	Y	Non-embedding
Hu et al. (2024)	36		40M-2B	400M-120B tokens	Y	Y	Non-embedding
Hashimoto (2021)				upto 600k sentences	Y	Y	NA
Ruan et al. (2024)	27* -77*		70B-180B	3T-6T tokens	N/A	N/A	N.S.
Anil et al. (2023)	12	1.00E+22	15B	4.00E+11	N	N	Non-embedding
Pearce & Song (2024a)	20 (simulated), 25 (real)		1.5B (simulated), 4.6M (real)	23B (simulated), 500M (real) tokens	Y	Y	w/ Embedding and Non-embedding considered separately
Cherti et al. (2023)	3* - 29		214M	34B (pretrain), 2B (finetune) examples	Y	Y	N.S
Porian et al. (2024)	16	2.00E+19	901M		Y	Y	w/ Embedding and Non-embedding considered separately
Alabdulmohsin et al. (2022)	1*		110M-1B	1e6-1e10 ex / 3e11 tokens	Mixed	N	N/A
Gao et al. (2024)	N.S	N.S	N.S	N.S	N	N	N.S
Muennighoff et al. (2024)	142		8,7B	900B tokens	Y	Y	w/ embedding
Rae et al. (2021)	4	6.31E+23	280B		Y	Y	Non-embedding
Shin et al. (2023)	17	0.1 PF Days	160M	500M-50B tokens	Y	Y	NA
Hernandez et al. (2022)	56		1.5M-800M	100B tokens	N	N	NS
Filipovich et al. (2022)	4		57-509M	30B token	Y	N	NS
Neumann & Gros (2022)	14		$5 * 10^5$	10^4 steps	Y	Y	NS
Droppo & Elibol (2021)	5-21		10^7	134-23k hrs speech	Y	Y	NS
Henighan et al. (2020)	6-10		10^11	10^12 tokens	Y	Y	Non-embedding
Goyal et al. (2024)	5		CLIP L/14 - 300M +63M	32-640M samples	Y	Y	Embedding
Aghajanyan et al. (2023)	21		8M-6.7B	5-100B tokens	Y	Y	Non-embedding
Kaplan et al. (2020)	40-150		1.5B	23B tokens	Y	Y	Non-embedding
Ghorbani et al. (2021)	12-14		191-3B	NS	Y	Y	Non-embedding
Gao et al. (2023)	9		3B	120-90k	N	Y	NS
Hilton et al. (2023)	NS	10^{20}			Y	Y	NS
Frantar et al. (2023)	48 and 112		0.66M-85M	1.8B images, 65B tokens	Y	Y	Non-embedding
Prato et al. (2021)	5			10^6 samples	Y	N	NA
Covert et al. (2024)	10		NA	1000 samples for IMDB	Y	Y	NA
Hernandez et al. (2021)	NS	10^{21}	10^8		Y	N	Non-embedding
Ivgi et al. (2022)	5-8		$10^4 - 10^8$	varies; 500k steps PT	Y	Y	Non-embedding
Tay et al. (2022a)			16-30B	2^19	Y	Y	NA
Tao et al. (2024)	60		33M-1.13B NV + 4-96k V	4.3B-509B Characters	Y	Y	Embedding and Non-embedding considered separately
Jones (2021)	200	1E+12-1E+17		4E+08-2E+09	Y	Y	NA
Zhai et al. (2022)	44		5.4M-1.8B	1-13M images	Y	Y	NA
Dettmers & Zettlemoyer (2023)	4		19M-176B	NA	NA	Y	NA
Dubey et al. (2024)	NS	$6 * 10^{18} - 10^22$	40M-16B		Y*	Y	NS
Hoffmann et al. (2022)	200-450	$6 * 10^{18} - 3 * 10^{21}$	16B	5B-400B tokens	Y	Y	Non-embedding
Ardalani et al. (2022)	NS	10^2-10^6 TFlops		5M-5B samples	N	N	All are considered
Clark et al. (2022)	56		15M-1.3B	130B tokens	Y	Y	Non-embedding

Table B.5: Details on training setup for each paper surveyed.

Paper	Data Points Per Law?	Scaling Law Metric	Modification Of Final Metric?	Subsets Of Data Used
Rosenfeld et al. (2019)	42-49	Loss / Top1 Error	N	N
Mikami et al.	7	Error Rate	N	N
Schaeffer et al. (2023)	NA	Various downstream	NA	NA
Sardana & Frankle (2023)	NS	Loss	NS	NS
Sorscher et al. (2022)	60	Error Rate	NA	NA
Caballero et al. (2022)	3-40	FID, Loss, Error Rate, Elo Score	N	NS
Besiroglu et al. (2024)	245	Loss	N	N
Gordon et al. (2021)	45-55	Loss	N	N
Bansal et al. (2022)	NS	Loss, BLEU	NS	NS
Hestness et al. (2017)	NS	Token Error, CER, Error Rate, Loss	Median min. validation error across multiple training runs with separate random seeds	NS
Bi et al. (2024)	upto 80	Validation bits-per-byte	NS	NS
Bahri et al. (2024)	upto 100	Loss	NS	NS
Geiping et al. (2022)	50	Effective Extra Samples	Interpolation	NS
Poli et al. (2024)	NS	Loss	NS	NS
Hu et al. (2024)	NS	Loss	NS	NS
Hashimoto (2021)	NS	Loss	NS	NS
Ruan et al. (2024)		Various downstream	N	N
Anil et al. (2023)	12	Loss	N	N
Pearce & Song (2024a)	20, 5	Loss	N	N
Cherti et al. (2023)	3-29	Error Rate	N	N
Porian et al. (2024)	12	Loss	N	N
Alabdulmohsin et al. (2022)	N.S.	Loss / Accuracy	N	N/A
Gao et al. (2024)	N.S	MSE	N.S	N.S
Muennighoff et al. (2024)	142	Loss	N	Outliers removed
Rae et al. (2021)	4	Loss	N/A	N/A
Shin et al. (2023)	NA	Loss	NA	NA
Hernandez et al. (2022)	NS	Loss	N	N
Filipovich et al. (2022)	NS	Loss	N	N
Neumann & Gros (2022)	238	Elo Score	N	N
Droppo & Elibol (2021)	NS	Loss	N	N
Henighan et al. (2020)	NS	Loss, Error Rate	NS	Drop smaller models
Goyal et al. (2024)	NS	Error Rate	N	N
Aghajanyan et al. (2023)	NS	Perplexity	N	N
Kaplan et al. (2020)	NS	Loss	NS	NS
Ghorbani et al. (2021)	NS	Loss, BLEU	Median of last 50k steps	
Gao et al. (2023)	90	RM Score	NS	NS
Hilton et al. (2023)	NS	Intrinsic Performance	Smoothing learning curve	Exclude early checkpoints
Frantar et al. (2023)	48 and 112	Loss	NS	NS
Prato et al. (2021)	5	Error Rate	NS	NS
Covert et al. (2024)	(1000-5000)*10	Expectation	NS	N
Hernandez et al. (2021)	40-120	Loss	NS	NS
Ivgi et al. (2022)	5-8	Loss	N	[2.5, 97.5] percentile
Tay et al. (2022a)	NA	Loss, Accuracy	NA	NA
Tao et al. (2024)	20*60	Loss	Interpolation	NS
Jones (2021)	2800	Elo Score	NS	NS
Zhai et al. (2022)	NS	Accuracy	NS	NS
Dettmers & Zettlemoyer (2023)	NA	Accuracy	NA	NA
Dubey et al. (2024)	150	Loss, Accuracy	NS	NS
Hoffmann et al. (2022)	upto 1500	Loss	N	Lowest loss model of a FLOP count, last 15% of checkpoints
Ardalani et al. (2022)	130	Loss	NS	NS
Clark et al. (2022)	26*56	Loss	Log	NS

Table B.6: Details on data extraction for each paper surveyed.

Paper	Curve-Fitting Method	Loss Objective	Hyperparameters Reported?	Initialization	Are Scaling Laws Validated?
Rosenfeld et al. (2019)	Least Squares Regression	Custom error term	N/A	Random	Y
Mikami et al.	Non-linear Least Squares in log-log space		N/A	N/A	Y
Schaeffer et al. (2023)	NA	NA	NA	NA	NA
Sardana & Frankle (2023)	L-BFGS	Huber Loss	Y	Grid Search	N
Sorscher et al. (2022)	NA	NA	NA	NA	NA
Caballero et al. (2022)	Least Squares Regression	MSLE	N/A	Grid Search, optimize one	Y
Besiroglu et al. (2024)	L-BFGS	Huber Loss	Y	Grid Search	Y
Gordon et al. (2021)	Least Squares Regression		N/A	N.S.	N
Bansal et al. (2022)	NS	NS	N	NS	N
Hestness et al. (2017)	NS	RMSE	N	NS	Y
Bi et al. (2024)	NS	NS	N	NS	Y
Bahri et al. (2024)	NS	NS	N	NS	N
Geiping et al. (2022)	Non-linear Least Squares		NA	Non-augmented parameters	Y
Poli et al. (2024)	NS	NS	N	NS	N
Hu et al. (2024)	scipy curvefit	NS	N	NS	N
Hashimoto (2021)	Adagrad	Custom Loss	Y	Xavier	Y
Ruan et al. (2024)	Linear Least Squares	Various	N/A	N/A	Y
Anil et al. (2023)	Polynomial Regression (Quadratic)	N.S.	N	N.S.	Y
Pearce & Song (2024a)	Polynomial Least Squares	MSE on Log-loss	N/A	N/A	N
Cherti et al. (2023)	Linear Least Squares	MSE	N/A	N/A	N
Porian et al. (2024)	Weighted Linear Regression	weighted SE on Log-loss	N/A	N/A	Y
Alabdulmohsin et al. (2022)	Least Squares Regression	MSE	Y	N.S.	Y
Gao et al. (2024)	N.S	N.S	N.S	N.S	N.S
Muennighoff et al. (2024)	L-BFGS	Huber on Log-loss	Y	Grid Search, optimize all	Y
Rae et al. (2021)	None	None	N/A	N/A	N
Shin et al. (2023)	NA	NA	NA	NA	NA
Hernandez et al. (2022)	NS	NS	NS	NS	NS
Filipovich et al. (2022)	NS	NS	NS	NS	NS
Neumann & Gros (2022)	NS	NS	NS	NS	NS
Droppo & Elibol (2021)	NS	NS	NS	NS	NS
Henighan et al. (2020)	NS	NS	NS	NS	NS
Goyal et al. (2024)	Grid Search	L2 error	Y	NA	Y
Aghajanyan et al. (2023)	L-BFGS	Huber on Log-loss	Y	Grid Search, optimize all	Y
Kaplan et al. (2020)	NS	NS	NS	NS	N
Ghorbani et al. (2021)	Trust Region Reflective algorithm, Least Squares	Soft-L1 Loss	Y	Fixed	Y
Gao et al. (2023)	NS	NS	NS	NS	Y
Hilton et al. (2023)	CMA-ES+Linear Regression	L2 log loss	Y	Fixed	Y
Frantar et al. (2023)	BFGS	Huber on Log-loss	Y	N Random Trials	Y
Prato et al. (2021)	NS	NS	NS	NS	NS
Covert et al. (2024)	Adam	Custom Loss	Y	NS	Y
Hernandez et al. (2021)	NS	NS	NS	NS	Y
Ivgi et al. (2022)	Linear Least Squares in Log-Log space	MSE	NA	NS	Y
Tay et al. (2022a)	NA	NA	NA	NA	NA
Tao et al. (2024)	L-BFGS, Least Squares	Huber on Log-loss	Y	N Random Trials from Grid	Y
Jones (2021)	L-BFGS	NS	NS	NS	NS
Zhai et al. (2022)	NS	NS	NS	NS	NS
Dettmers & Zettlemoyer (2023)	NA	NA	NA	NA	NA
Dubey et al. (2024)	NS	NS	NS	NS	Y
Hoffmann et al. (2022)	L-BFGS	Huber on Log-loss	Y	Grid Search, optimize all	Y
Ardalani et al. (2022)	NS	NS	NS	NS	NS
Clark et al. (2022)	L-BFGS-B	L2 Loss	Y	Fixed	NS

Table B.7: Details on optimization for each paper surveyed.

B.4 Recommendations

As seen in our analyses, many decisions in our checklist have a number of reasonable options, but those reasonable choices lead to a wide range of scaling law fits, and the observed variations do not follow any clear pattern. It is probable that variations would be even harder to predict when varying model architectures or other design decisions, removing the possibility of a universal set of best practices. However, it is certainly possible to determine that some scaling law fits are plausible or highly implausible, and to observe the stability of the fitting procedure. With the caveat that following any recommendations can not guarantee good scaling law fit, we can make some more concrete recommendations based on these observations:

Scaling Law Hypothesis

- Fitting fewer scaling law parameters at a time typically results in greater stability. In some cases, it may be beneficial to decompose the scaling law fitting problem into two separate procedures. Examples of this approach are the IsoFLOP procedure from Hoffmann et al. (2022), as well as fitting first the relation between L and C, then finding the optimal N and D for a C, as seen in Porian et al. (2024).

Training Setup

- The trained models should include a wide range of input variables settings. For example, when the input variables to the scaling law are L, N, D, the included models should include a wide range of N and D values for each C, or equivalently, should include a wide range of D/N ratios. If the included settings do not include the true optimum, the procedure will struggle to fit to the optimum.

- Sweeping for the optimal learning rate results in a less stable fit than fixing the learning rate. We hypothesize that this may be because the true optimal learning rate for each model and data budget size is not any of the options we consider, and thus, each model varies in the difference between its true and approximate optimal learning rate. This may introduce additional noise to the data. Due to resource constraints, we are unable to fully test this hypothesis, and it may not hold at significantly larger scale, but we recommend fixing the learning rate, or changing it according to, say, model size, according to a fixed formula.

Data Collection

- Results across tasks or datasets should not be mixed. Neither performance predictions nor optimal D/N ratios are fixed across different evaluation settings for the same set of models.

Fitting Algorithm

- Scaling law fitting is sensitive to initialization; most known optimization methods for scaling laws are only able, in practice, to the shift parameters near to their initialization values. Thus, a dense search over initialization values is necessary. If there is a strong hypothesis guiding the choice of one specific initialization, such as a previously fit and validated scaling law, this will also limit the set of possible final scaling law parameter values.
- Different losses emphasize the contribution of errors from certain datapoints. The chosen loss should be suited to the distribution of datapoints and sources of noise.
- A simple grid search is unlikely to result in a good scaling law fit. Additionally, optimizers designed to fit linear relations may make assumptions about the distribution of errors and should not be used to fit a power law.

B.4.1 Example Checklist

We provide one possible set of responses to our checklist, reflective of some recommendations enumerated above, and loosely based on Hoffmann et al. (2022). These answers roughly correspond to a subset of the experiments we run in §3.1.9.

(Mis)Fitting: Scaling Law Reproducibility Checklist

Scaling Law Hypothesis (§3.1.3)

- What is the form of the power law? $L(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta}$
- What are the variables related by (included in) the power law? *N: the number of model parameters, D: the number of data tokens, and L: the model's validation loss*
- What are the parameters to fit? *A, B, E, α, β*
- On what principles is this form derived? *This is taken from Hoffmann et al. (2022), who hypothesize this form on the basis of prior work in risk decomposition.*
- Does this form make assumptions about how the variables are related? *This form inherently assumes that N and D do not have any interaction in their effect on the scaling of L. For some experiments, we use the assumption $\alpha = \beta$ to simplify optimization.*

Training Setup (§3.1.6)

- How many models are trained? *Refer to Table 10*
- At which sizes? *Refer to Table 10*
- On how much data each? On what data? Is any data repeated within the training for a model? *Refer to Table 10.*
- How are model size, dataset size, and compute budget size counted? For example, how are parameters of the model counted? Are any parameters excluded (e.g., embedding layers)? How is compute cost calculated? *We include the results including and excluding embedding layers for both the total parameter count N and the total FLOP count C. We also include, for comparison, results using the estimate $C = 6ND$.*
- Are code/code snippets provided for calculating these variables if applicable? `https://github.com/hadasah/scaling_laws`

- How are hyperparameters chosen (e.g., optimizer, learning rate schedule, batch size)? Do they change with scale? *Most hyperparameters are chosen based on best practices in current literature; several are taken directly from the settings in Hoffmann et al. (2022). For learning rate, we conduct an extensive hyperparameter search across 2-3 orders of magnitude, multiplying by 2-2.5, and then conduct training at 3 learning rates, including the optimum, for nearly all (N, D) configurations.*
- What other settings must be decided (e.g., model width vs. depth)? Do they change with scale? *Refer to Table 10*
- Is the training code open source? *Yes*

Model size	# layers	Hidden Dim	# Attn Heads	# Steps
12M	5	448	7	{100, 200, 250, 360, 500, 750, 1,000, 4,000}
17M	7	448	7	{100, 200, 250, 500, 750, 1,000, 1,250, 1,500}
25M	8	512	8	{250, 360, 500, 750, 1,000, 1,500, 2,000, 8,000, 16,000}
35M	9	576	9	{200, 250, 360, 500, 750, 1,000, 1,250, 1,500, 4,000, 16,000}
50M	10	640	10	{250, 500, 750, 1,000, 1,250, 1,500, 1,800, 2,000, 2,500, 4,000, 16,000}
70M	12	704	11	{250, 500, 750, 1,000, 1,250, 1,500, 2,000, 2,500, 3,000, 5,000}
100M	14	768	12	{250, 500, 1,000, 1,500, 2,000, 2,500, 3,000, 4,000, 6,000, 12,000}
200M	18	960	15	{500, 1,500, 6,000, 9,000}
300M	19	1152	18	{4,000}
400M	20	1280	20	{3,000, 6,000}
1B	26	1792	28	{20,000}

Table B.8: Architecture and Data budget details for models trained in §3.1.9

Data Collection(§3.1.7)

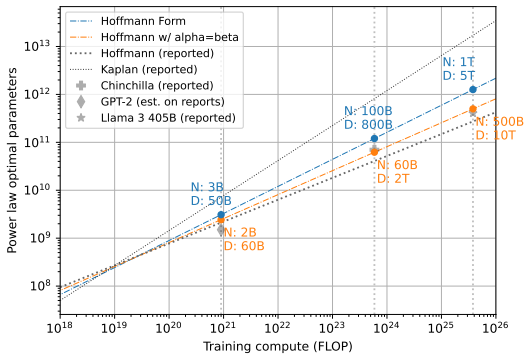
- Are the model checkpoints provided openly? *Yes, at* `https://github.com/hadasah/scaling_laws`
- How many checkpoints per model are evaluated to fit each scaling law? Which ones, if so? *Unless clearly denoted otherwise, one checkpoint per model is evaluated; the last checkpoint. By default, no mid-training checkpoints are used, i.e., from before the termination of the cosine learning rate schedules.*
- What evaluation metric is used? On what dataset? *We use cross-entropy loss, measured on a held-out validation subset of the Common Crawl (Raffel et al., 2020) dataset.*
- Are the raw evaluation metrics modified? Some examples include loss interpolation, centering around a mean or scaling logarithmically. *No.*
- If the above is done, is code for modifying the metric provided? *Yes.*

Fitting Algorithm (§3.1.8)

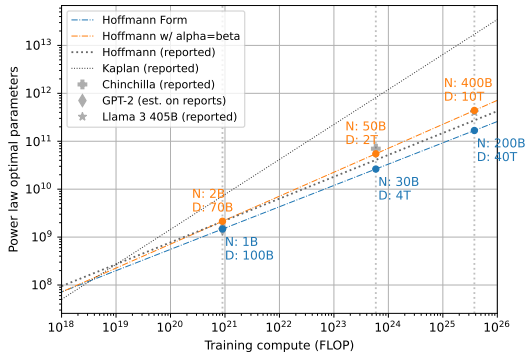
- What objective (loss) is used? *We try various loss objectives (1) log-Huber loss, (2) MSE, (3) MAE and (4) Huber loss*
- What algorithm is used to fit the equation? *Mainly L-BFGS, but we also experiment with BFGS, non-linear least squares and grid search.*
- What hyperparameters are used for this algorithm? *Thresholds of* $\{1e-4, 1e-6, 1e-8\}$*, exact gradient*
- How is this algorithm initialized? *We initialize with 4500 initializations similar to Hoffmann et al. (2022).*
- Are all datapoints collected used to fit the equations? For example, are any outliers dropped? Are portions of the datapoints used to fit different equations? *No outliers are dropped in general, but we do show some results on specific subsets of models. For example, we compare the result of a scaling law fit when using only models trained at a peak learning rate of* $1e-4$ *or* $4e-4$*.*

- How is the correctness of the scaling law considered? Extrapolation, Confidence Intervals, Goodness of Fit? *Currently, we do not evaluate the correctness beyond comparing to results in literature (Hoffmann et al., 2022; Kaplan et al., 2020).*

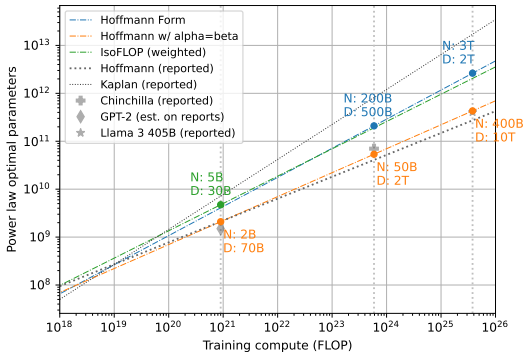
B.5 Full Analysis Plots



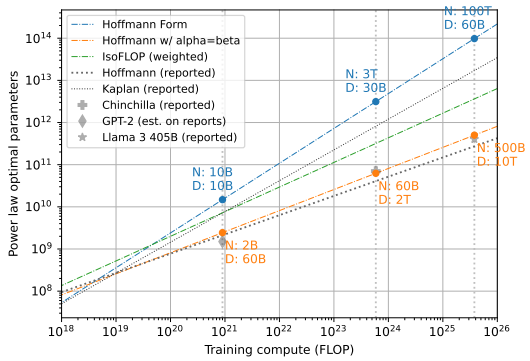
Hoffmann et al. (2022); Besiroglu et al. (2024)



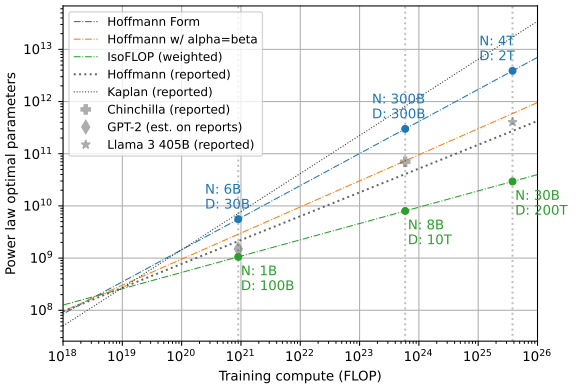
Porian et al. (2024)



Porian et al. (2024) (all checkpoints)

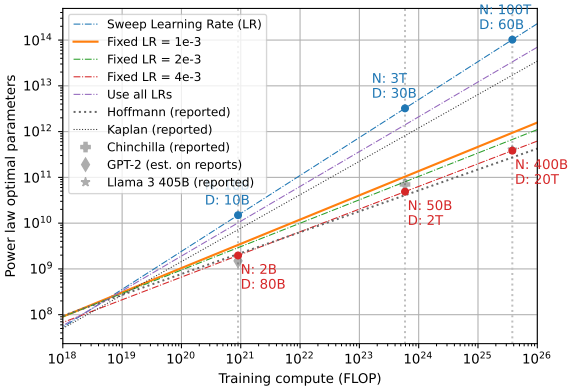


Ours

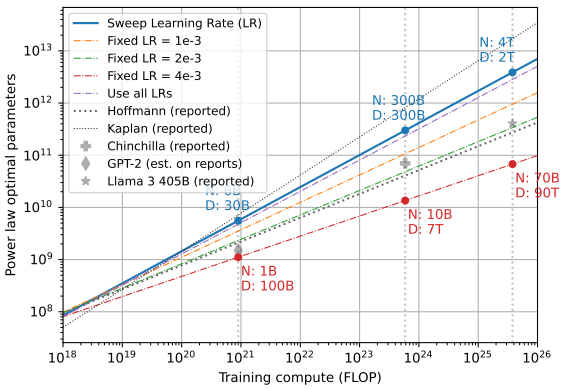


Ours (all checkpoints)

(a) **§3.1.3, §3.1.10** Using data from both Besiroglu et al. (2024) (left) and our own models (right), we compare the effects of fitting to the power law form used in Approach 3 of Hoffmann et al. (2022) with the variant used by Muennighoff et al. (2024), which assumes that the exponents α, β are equal – equivalently, that $N^*(C)$ and $D^*(C)$ scale about linearly with each other. When using only the performance of final checkpoints from both Besiroglu et al. (2024) and our own experiments, taking this assertion results in a law much closer to the one reported by Hoffmann et al. (2022). On our own models, we also show results when using the IsoFLOP approach from Hoffmann et al. (2022). As we are using only results from final model checkpoints, the size of the data input to the IsoFLOP approach in this case is reduced.

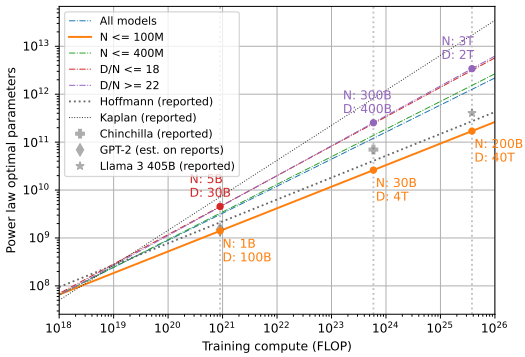


Ours

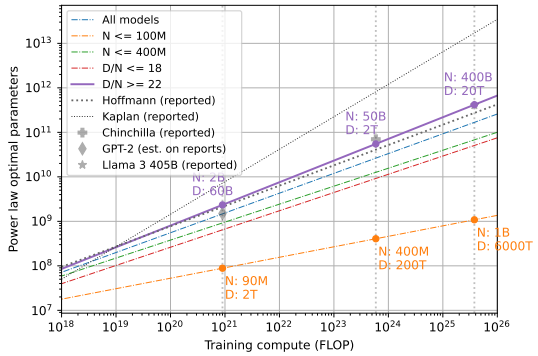


Ours (all checkpoints)

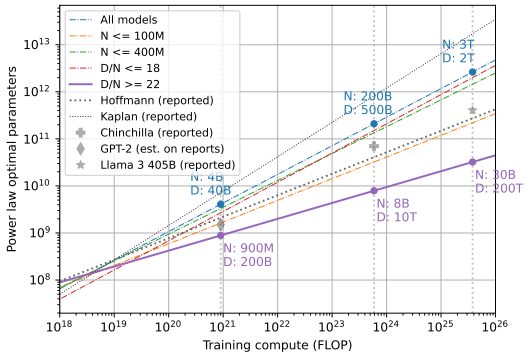
(b) **§3.1.6, §3.1.11** With our models, we simulate the effects of not sweeping the learning rate. As a baseline, (1) we sweep at each (N, D) pair for the optimal learning rate over a range of values at most a multiple of 2 apart. Next, (2) we use a learning rate of 1e-3 for all N, the optimal for our 1 billion parameter models, and do the same for (3) 2e-3 and (4) 4e-3, which is optimal for our 12 million parameter models. Lastly, we use all models across all learning rates at the same N and D. Results vary dramatically across these settings. Somewhat surprisingly, using all learning rates results in a very similar power law to sweeping the learning rate, whereas using a fixed learning rate of 1e-3 or 4e-3 yields the lowest optimization loss or closest match to the Hoffmann et al. (2022) power laws, respectively.



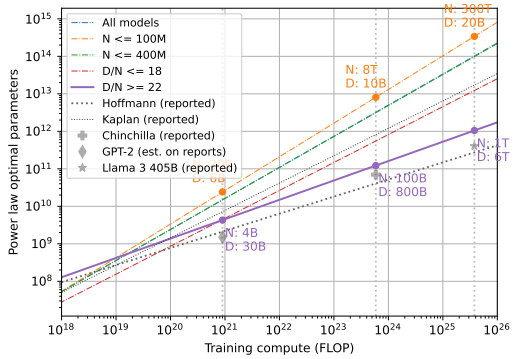
Hoffmann et al. (2022); Besiroglu et al. (2024)



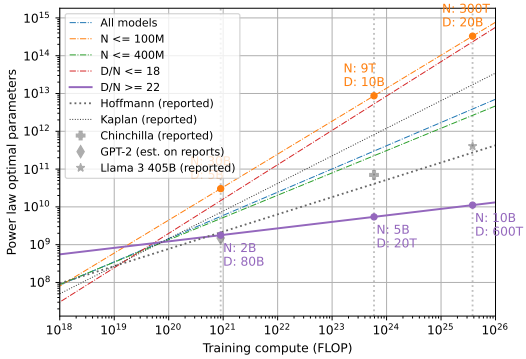
Porian et al. (2024)



Porian et al. (2024) (all checkpoints)

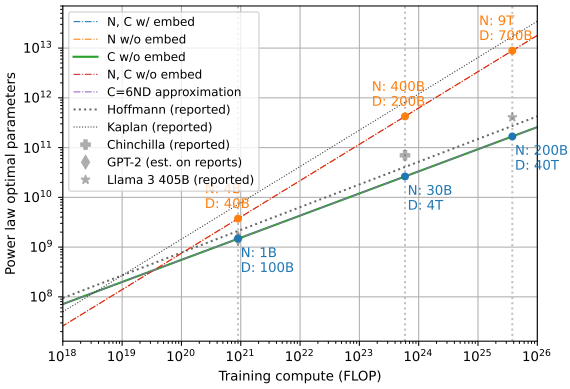


Ours

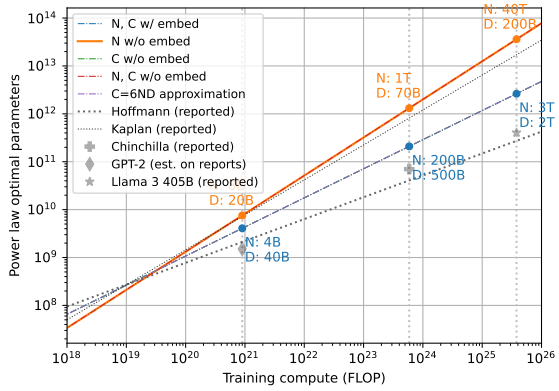


Ours (all checkpoints)

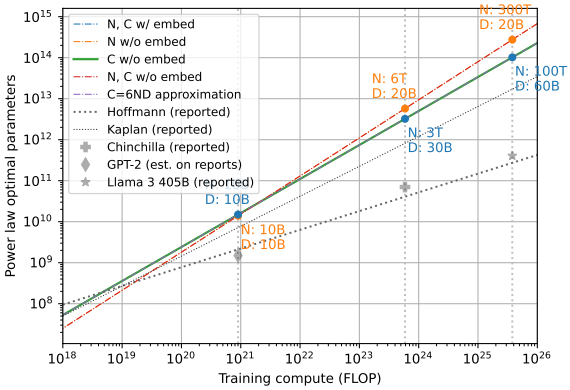
(c) **§3.1.6, §3.1.11** From all 3 datasets, we choose subsets with (N, D) values which fit a particular method one might have of setting up training. We fit with (1) all models, which for our dataset, ranges from 12 million to 1 billion parameters, then with (2) only models of up to about 100 million parameters or (3) up to 400 million parameters. We also compare the effects of a higher or lower hypothesis about the optimal D/N ratio, including (4) only models with $D/N \leq 18$ or (5) $D/N \geq 22$. These ranges are designed to exclude $D/N = 20$, the rule of thumb based on Hoffmann et al. (2022). The minimum or maximum D/N ratio tested does skew results; above 10^{22} FLOPs, (4) and (5) fit to optimal ratios $D/N < 18$ and $D/N > 22$, respectively. Removing our largest models in (2) also creates a major shift in the predicted optimal D/N.



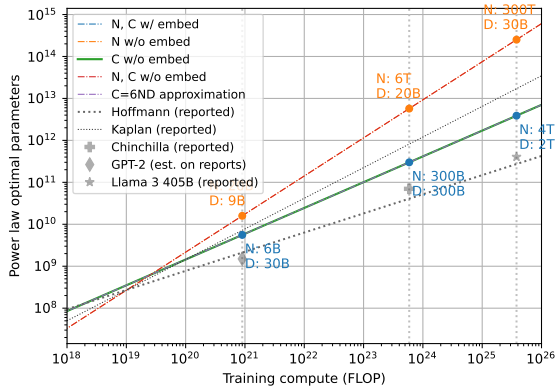
Porian et al. (2024)



Porian et al. (2024) (all checkpoints)

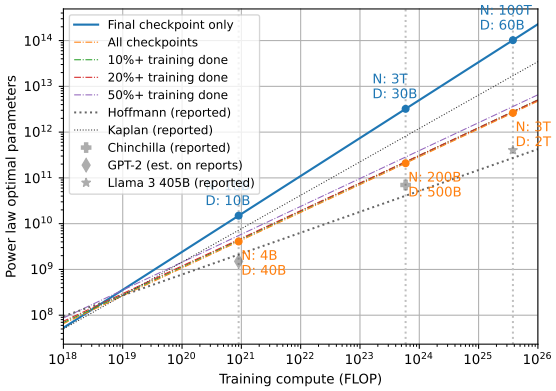


Ours

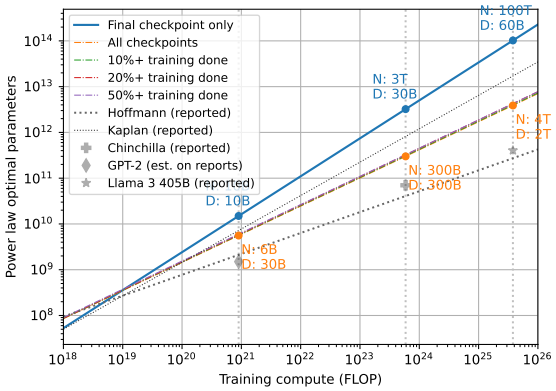


Ours (all checkpoints)

(d) **§3.1.6, §3.1.11** With our own data and those from Porian et al. (2024), we fit power laws to the same sets of models, while varying the ways we count N and C. We compare (1) including embeddings, which is our baseline, with (2) excluding embeddings only in N, (3) excluding embeddings only in C, (4) excluding embeddings in both N and C. We also compare to using the $C = 6ND$ approximation, including embedding parameters. Throughout this work, we calculate the FLOPs in a manner similar to Hoffmann et al. (2022), and we open source the code for these calculations. With both datasets, the exclusion of embeddings in FLOPs has very little impact on the final fit. Similarly, using the $C = 6ND$ approximation has no visible impact. For the Porian et al. (2024) models, the exclusion of embedding parameters in the calculation of N results in scaling laws which differ substantially, and with increasing divergences at large scales.

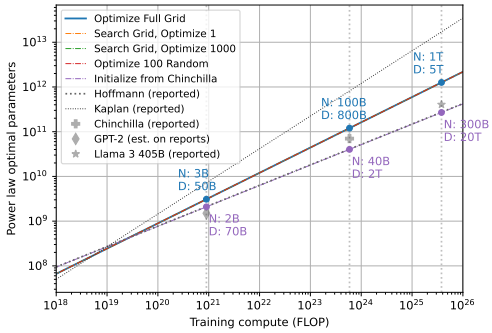


Porian et al. (2024) (all checkpoints)

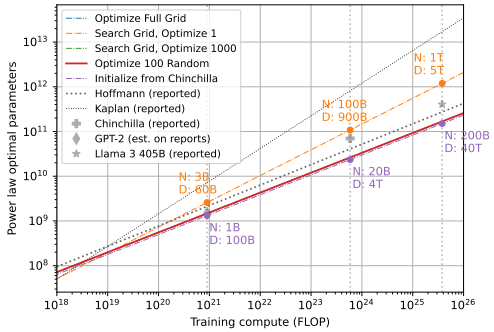


Ours (all checkpoints)

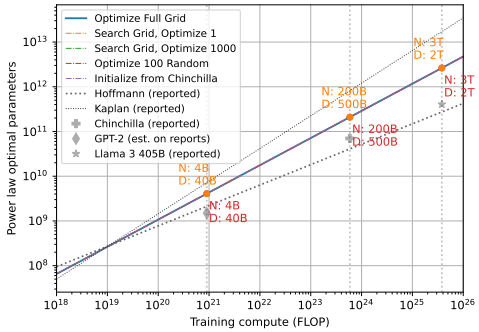
(e) **§3.1.7, §3.1.12** With models from Porian et al. (2024) and our own dataset, we compare (1) fitting a power law using only the final checkpoint with (2) using all mid-training checkpoints (3) using all checkpoints, starting 10% through training, (4) the same, starting 20% through training, and (5) the same again, starting 50% through training. We observe that (2)-(5) consistently yields power laws more similar to that reported by Hoffmann et al. (2022), so we also repeat all analyses in Figures B.1a-B.1d. Using mid-training checkpoints sometimes results in more stable fits which are similar to the Hoffmann et al. (2022) scaling laws, but the effect is noisy and dependent on other decisions



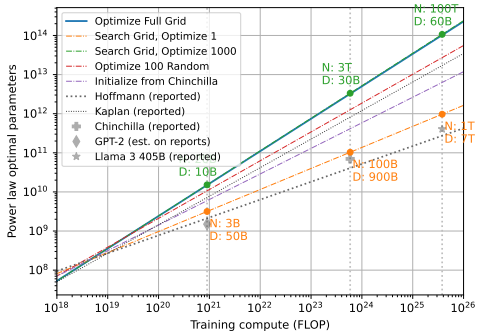
Hoffmann et al. (2022); Besiroglu et al. (2024)



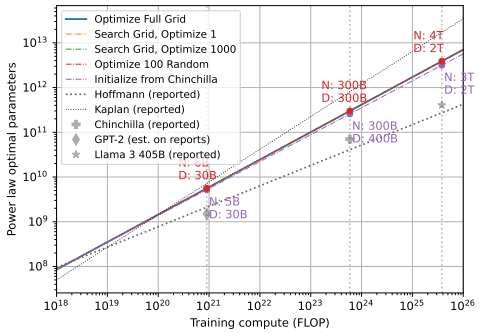
Porian et al. (2024)



Porian et al. (2024) (all checkpoints)

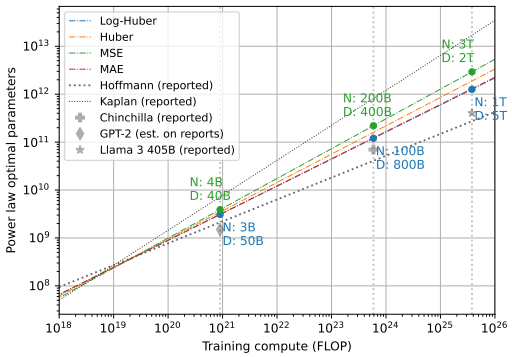


Ours

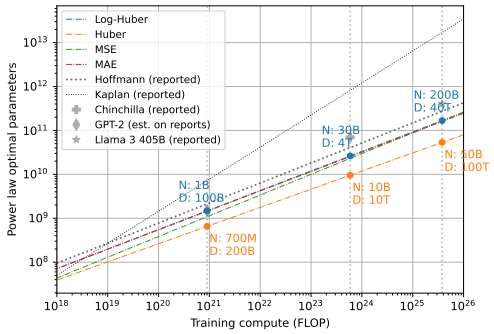


Ours (all checkpoints)

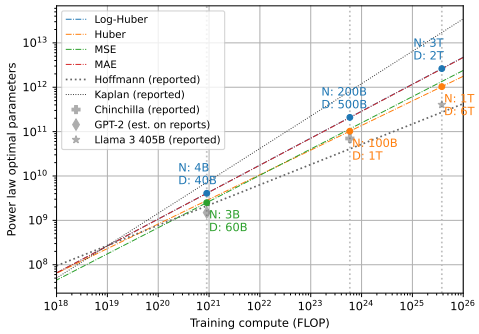
(f) **§3.1.8, §3.1.13** We fit to data from all 3 datasets to experiment with the initialization of parameters in the power law. We start with (1) optimizing every point in a grid search of 6x6x5x5x5=4500 initializations (Hoffmann et al., 2022), (2) randomly sampling from only a single initialization in this grid, (3) searching for the lowest loss initialization point (Caballero et al., 2022), (4) randomly sampling 100 points, and (5) initializing with the coefficients found in Hoffmann et al. (2022), as Besiroglu et al. (2024) does. With the Besiroglu et al. (2024) data, (5) yields a fit nearly identical to that reported by Hoffmann et al. (2022), although (1) results in the lowest fitting loss. With the Porian et al. (2024) data, all approaches except (2) yield a very similar fit, which gives a recommended D/N ratio similar to that ofHoffmann et al. (2022). However, using our data, (2) optimizing over only the most optimal initialization yields the best match to the Hoffmann et al. (2022) power laws, followed by (5) initialization from the reported Hoffmann et al. (2022) scaling law parameters. Optimizing over the full grid yields the power law which diverges most from the Hoffmann et al. (2022) law, suggesting the difficulty of optimizing over this space, and the presence of many local minima.



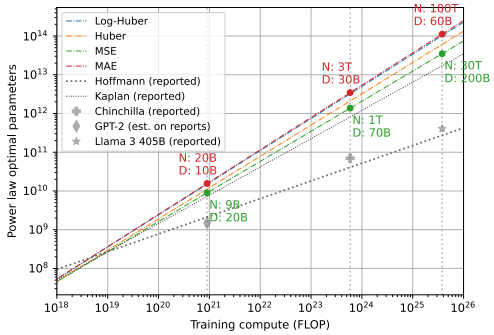
Hoffmann et al. (2022); Besiroglu et al. (2024)



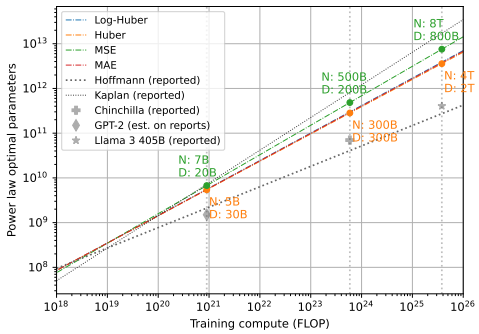
Porian et al. (2024)



Porian et al. (2024) (all checkpoints)

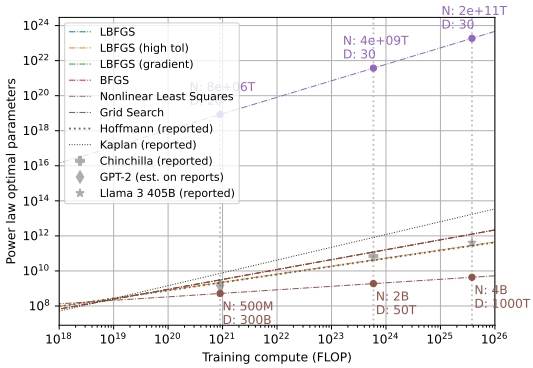


Ours

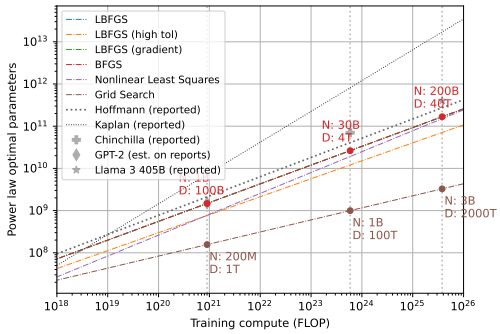


Ours (all checkpoints)

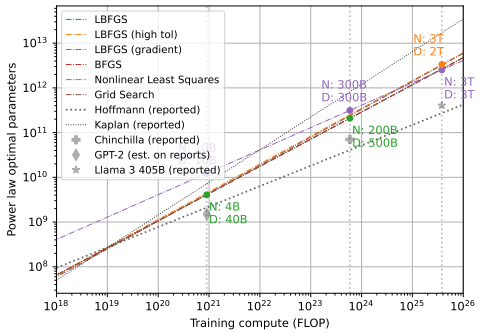
(g) **§3.1.8, §3.1.13** We fit a power law to data from from all 3 datasets, minimizing different objective functions: (1) the baseline log-Huber loss, (2) MSE, (3) MAE, and (4) the Huber loss. The resulting power laws are less disparate than when varying many of the other factors discussed above and generally fall near the power law parameters reported by Kaplan et al. (2020) and Hoffmann et al. (2022), but this is still a wide range of recommended optimal parameter counts for each compute budget. Overall, the loss function behavior is not predictable, given the differences between loss functions when looking at the power laws resulting from these three sources of data.



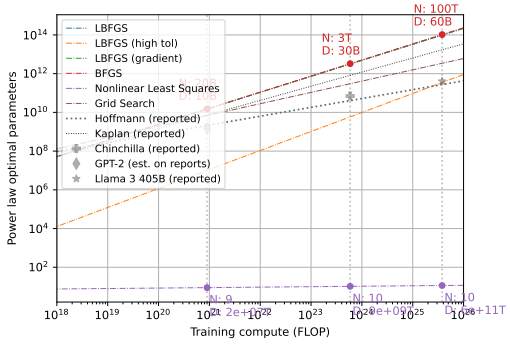
Hoffmann et al. (2022); Besiroglu et al. (2024)



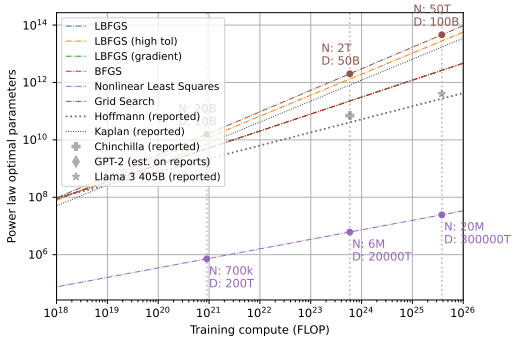
Porian et al. (2024)



Porian et al. (2024) (all checkpoints)



Ours



Ours (all checkpoints)

(h) **§3.1.8, §3.1.13** We fit a power law to data from all 3 datasets using various optimizers, beginning with the original (1) L-BFGS. L-BFGS and BFGS implementations have an early stopping mechanism, which conditions on the stability of the solution between optimization steps. We set this threshold for L-BFGS to a (2) higher value 1e-4 (stopping earlier). We found that using any higher or lower values resulted in the same solutions or instability for all three datasets. L-BFGS and BFGS also have the option to use the true gradient of the loss, instead of an estimate, which is the default. In this figure, we include this setting, (3) using the true gradient for L-BFGS. We compare these L-BFGS settings to (4) BFGS. We test the same tolerance value and gradient settings, and find that none of these options change the outcome of BFGS in our analysis, and omit them from this figure for legibility. Finally, we compare to (5) non-linear least squares and (6) pure grid search, using a grid 5 times more dense along each axis as we used initialization with other optimizers. This density is chosen to approximately match the runtime of L-BFGS. Many of these optimizers do converge to similar solutions, but this depends on the data, and the settings which diverge from this majority do not follow any immediately evident pattern.

Figure B.1: (**§3.1.9**) We study the effects of various decisions in the fitting of a power law, as outlined in our checklist (Appendix B.2) and detailed in §3.1.3-§3.1.8. For each set of analyses, we the scaling laws found by (Kaplan et al., 2020) and (Hoffmann et al., 2022) for comparison. We also include markers indicating 3 existing models for comparison purposes: Llama 3 405B (Dubey et al., 2024), the Chinchilla model (Hoffmann et al., 2022), and an estimate of the 1.5B GPT-2 model (Radford et al., 2019), for which we know details of the dataset storage size and word count, but not an exact count of data BPE tokens, which we estimate at 100B. We additionally annotate, at the compute budget C for each of these 3 reference points, the maximum and minimum *predicted* (i.e. extrapolated) optimal model parameter count N_{opt} and data budget D_{opt} from the fitted power laws. We use a thicker, solid line for the method in each plot which achieves the lowest optimization loss, with the exception of the plots comparing power law form, those comparing loss functions and those comparing optimizers, for which this would be nonsensical. We find overall, throughout our analyses, that all of the decisions we explore have an impact on the final fit of the power law, supporting our conclusion that more thorough reporting of these decisions is critical for scaling law reproducibility.

Appendix C

ATLAS: SUPPLEMENTARY MATERIALS

C.1 Extended Related Work

Scaling Laws Scaling laws are used to study the behavior of deep learning systems on scaling training data and/or compute. Various researchers have observed scaling properties of generalization error with training data size and model capacity (Banko & Brill, 2001; Hestness et al., 2017; Amodei et al., 2016). Specifically, in the context of LLMs, Kaplan et al. (2020) explicitly propose a power law for the relationship between the loss L, number of parameters in the language model N, and the number of dataset tokens D. Later, Hoffmann et al. (2022) proposed a different formula:

$$L(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} \tag{C.1}$$

We build upon this form throughout the paper to fit our scaling laws. Other researchers, too, have built upon these scaling laws to study various aspects of scaling, such as neural network architectures (Clark et al., 2022; Tay et al., 2022a; Frantar et al., 2023; Gu & Dao, 2023; Scao et al., 2022) or transfer learning (Henighan et al., 2020). Researchers have also investigated scaling properties in different domains of deep learning such as neural machine translation (Ghorbani et al., 2021; Gordon et al., 2021), vision language models like CLIP (Cherti et al., 2023; Henighan et al., 2020), vision (Alabdulmohsin et al., 2022; Zhai et al., 2022), reinforcement learning (Hilton et al., 2023; Jones, 2021; Gao et al., 2023), and recommendation systems (Ardalani et al., 2022).

Scaling Laws for Data Mixtures Many prior works (Hoffmann et al., 2022) assume a fixed dataset distribution to study the effect of scaling on model performance. However, empirically, dataset composition can have a significant impact on quality (Longpre et al., 2023b; Albalak et al., 2024; Sorscher et al., 2022). Hashimoto (2021) studies the relationship between dataset composition and model loss, finding a simple scaling law quantifying this relationship. Bansal et al. (2022) study the impact of data quality and noise

on architecture, finding that synthetic data has a different exponent. Fernandes et al. (2023) study scaling for multilingual neural translation (Bapna et al., 2022), but experiments are restricted to 2-3 language pairs. Aghajanyan et al. (2023), in a similar vein, propose bimodal scaling laws for multimodal foundation models (Reid et al., 2024). Other researchers have considered different aspects of dataset composition that can affect model scaling: Muennighoff et al. (2024) study the effect of data repetition on scaling, while Goyal et al. (2024) consider how the effect of mixing data pools of varying quality changes with scale.

Foundation models involve training models on a mixture of datasets with the aim of achieving a certain validation loss on various validation sets and downstream datasets of interest. Based on the observation that optimal data mixture changes with scale, several recent works have attempted to characterize the change in the scaling law equation when either the train or test distribution changes.

Some researchers predict the optimal data mixture by first ablating mixtures using smaller models, and second, training a larger scaled-up model using the found optimal mixture (Ye et al., 2024; Xie et al., 2024; Liu et al., 2024). However, it is unclear if this optimal data mixture can extrapolate to a much larger magnitude, with some evidence in computer vision that this is not the case (Goyal et al., 2024) - i.e., the optimal data mixture changes with scale.

Some researchers have described scaling laws for different downstream evaluation sets for the same model (Dubey et al., 2024; Chen et al., 2024), while others (Brandfonbrener et al., 2024) have derived power law relationships between models trained on different distributions and the same model on different test sets.

Recently, several researchers have attempted to predict the final loss of a model for a given data mixture with scaling laws. Ge et al. (2024b) use the observation of a logarithmic shift in scaling pattern to predict loss on subsets of the Pile weighted in various proportions on a fixed model size. Ye et al. (2024), too, fit a scaling law for data mixtures based on a single scale. Kang et al. (2024) empirically show that the optimal mixture for a model changes with scale, and propose a method that addresses this on GPT-2 scale models. Que et al. (2024) use continual pre-training as a tool to develop their scaling law. Shukor et al. (2025), too, develop scaling laws that account for the transfer between data mixtures changing at scale for English LLMs and multimodal models for less than 10 domains.

Multilingual Scaling Laws Prior work has also investigated multilingual transfer effects specifically in pretraining (Chang et al., 2024) and post-training (Shimabucoro et al., 2025), though on a smaller scale of models. He et al. (2024) is the work closest to ours, where the authors attempt to describe scaling laws for

multilingual LLMs. A crucial distinction from our work is that we account for individual-language transfer learning in our scaling laws, and achieve a better resulting fit (Table 4.1).

C.2 Experimental Details

To understand multilingual scaling characteristics we run 774 separate experiments, across different model scales, and language mixtures. In Table C.1 we summarize the experiment segments, using the notation below, that defines sets of languages or scales.

- $\mathcal{L}_{\text{mono}}$ (7): {en, fr, ru, zh, hi, sw, yo}
- $\mathcal{L}_{\text{pairs}}$ (10): {en, fr, ru, zh, hi, de, es, pt, ja, vi}
- $\mathcal{L}_{\text{target}}$ (15): en/X for $X \in \mathcal{L}_{\text{pairs}}$ + {es/pt, fr/zh, hi/ru, de/ja, hi/vi}
- $\mathcal{L}_{\text{eval}}$ (48): complete evaluation set (all 48 languages)
- $\mathcal{C}_{\text{exp}}$ (12): {4×4, 6, 8×2, 12, 16, 24, 32, 50}
- $\mathcal{S}_{\text{full}}$ (20): {0, 1, 2, 4, 6, 8, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 42, 46}
- $\mathcal{S}_{\text{part}}$ (11): {0, 4, 8, 12, 16, 20, 24, 28, 34, 42, 46}
- $\mathcal{S}_{\text{min}}$ (7): {0, 8, 16, 24, 34, 42, 46}

To see the model sizes defined by the scales in $\mathcal{S}_{\text{full}}$, $\mathcal{S}_{\text{part}}$, and $\mathcal{S}_{\text{min}}$, see Table C.3. We select these scale ranges to empirically observe models from $10M$ to $8B$ parameters. Subsets of languages, or subsets of the full scale ($\mathcal{S}_{\text{part}}$, or $\mathcal{S}_{\text{min}}$) are chosen to ensure the experiment number is sufficiently tractable for each research question, but also computationally feasible given our resources.

C.2.1 Language Choice

The fifty languages we selected come from the intersection of languages covered by both Flores-200 and MADLAD-400. They were selected to cover a wide range of language families, scripts, and resource levels. To get this collection, we sorted all languages in the Flores-MADLAD intersection by the number of characters, and then selected every six languages. We then went over this list and perturbed the index up

or down if it was too typologically similar to an already-selected language. For example, Language 48 was Afrikaans, but there were already several Germanic languages on the list, so we opted to take language 47 instead (Telugu).

C.2.2 Experiment Design

Pretraining Hyperparameter Details We rely on the hyperparameter settings refined by prior work to obtain reliable performance across model sizes. We also conducted initial experiments, varying batch sizes, learning rates, optimizer, and dropout to ensure our settings were relatively robust. In table C.2 we enumerate and explain all our hyperparameter choices, citing related work.

Train & Test Data We use two test sets. For each we measure the vocabulary-insensitive loss proposed in Tao et al. (2024) in order to fairly evaluate across languages.

1. **MADLAD-400 Test** We isolate a test set from each of our 50 evaluation languages in MADLAD-400 (Kudugunta et al., 2024). In initial experiments we find stable metrics over random seeds requires sampling roughly (sequence length x num instances) $2048 * 10000 = 20{,}480{,}000$ tokens. For each language we randomly sample the minimum of either 20M tokens or 20% of the total tokens for low resource languages min(20M tokens, 20%).

2. **Flores-101** We use Flores-101 (Goyal et al., 2021), a language-parallel set of 101 high-quality translations, as a comparable test set. We extract the monolingual sequences the test set, using those to compute sequence losses, comparable to MADLAD-400 Test.

Vocabulary Details We use a vocabulary trained by Kudugunta et al., with a 64000 sized vocabulary, $T = 100$ and 99.9995% character coverage.

Table C.2: **The pretraining hyperparameters used across experiments.** We detail the batch size scheduling, the vocabulary choice, as well as fine-grained hyperparameter choices. Each choice is justified and grounded in prior work, discussed in the Explanation column.

PARAMETER	VALUE	EXPLANATION
General Training Parameters		
Learning Rate Scheduler	WSD	We adopt the Warmup Stable Decay (WSD) learning rate schedule, designed to efficiently study data-model scaling law without extensive retraining (Hu et al., 2024; Hägele et al., 2024)
Optimizer	AdamW	Commonly used in scaling law experiments (Loshchilov & Hutter, 2019; Hoffmann et al., 2022)
Base Learning Rate	2e-4	Following (Muennighoff et al., 2024) and confirmation it works well in initial experiments.
Warmup Steps	1000	Standard practice, as in (Longpre et al., 2023b)
Decay Period	10%	Following (Hu et al., 2024; Hägele et al., 2024) this is the minimum well performing decay length.
Sequence Length	2048	As in (Longpre et al., 2023b; Muennighoff et al., 2024).
Training Steps	30k+	We vary the steps according to experiments, to always ensure we are well above likely chinchilla compute optimal ranges.
Dropout	0.1	While English scaling laws work tends to use a dropout of 0.0, we see little notable difference, except for lower resource languages, with repeating epochs.
Batch Size Schedule		

Continued on next page

Table C.2 continued from previous page

PARAMETER	VALUE	EXPLANATION
Batch Size (<150M params)	256	Hu et al. (2024); Muennighoff et al. (2024) both find
Batch Size (<1B params)	512	the optimal batch size increases with model size. We
Batch Size (<2B params)	1024	adopt a slightly larger rising batch size than
Batch Size (2B+ params)	2048	Muennighoff et al. (2024), following initial empirical results.
		Vocabulary
Vocab Size	64k	
Vocab (Unimax)		Vocabulary by Kudugunta et al., with $T = 100$ and 99.9995% character coverage.
Vocab (Monolingual)		Same settings as above, but trained only on sentences from the language being considered.

Table C.3: **The SCALE value mapped to the dimensions of the model, and the parameter count.** The model dimensions, borrowed from Muennighoff et al. (2024), report the number of attention heads, number of layers, embed size, feedforward size, and key-value size. Our model may be slightly different sizes than prior work based on our vocabulary size (64000 + 512 special tokens).

SCALE	HEADS	LAYERS	EMBED	FFW	KV	PARAMETERS
0	4	3	128	512	32	9,044,352
1	7	4	224	896	32	17,662,848
2	7	5	288	1,152	32	24,847,776
3	7	6	448	1,792	32	45,763,200
4	8	8	512	2,048	64	66,588,672
5	9	9	576	2,304	64	84,939,840
6	10	10	640	2,560	64	106,830,080
7	10	13	640	2,560	64	126,492,800
8	10	16	640	2,560	64	146,155,520
9	12	12	768	3,072	64	162,800,640
10	12	15	768	3,072	64	191,114,496

Continued on next page

Table C.3 continued from previous page

Scale	Heads	Layers	Embed	FFW	KV	Parameters
11	12	18	768	3,072	64	219,428,352
12	14	14	896	3,584	64	237,646,080
13	14	16	896	3,584	64	263,337,984
14	14	18	896	3,584	64	289,029,888
15	16	16	1,024	4,096	64	334,512,128
16	16	18	1,024	4,096	64	368,068,608
17	16	20	1,024	4,096	64	401,625,088
18	10	18	1,280	5,120	128	554,457,600
19	10	21	1,280	5,120	128	633,104,640
20	11	18	1,408	5,632	128	661,807,872
21	10	24	1,280	5,120	128	711,751,680
22	11	21	1,408	5,632	128	756,970,368
23	12	19	1,536	6,144	128	816,345,600
24	11	24	1,408	5,632	128	852,132,864
25	12	22	1,536	6,144	128	929,596,416
26	12	25	1,536	6,144	128	1,042,847,232
27	14	20	1,792	7,168	128	1,143,245,824
28	14	23	1,792	7,168	128	1,297,391,872
29	14	26	1,792	7,168	128	1,451,537,920
30	16	22	2,048	8,192	128	1,608,560,640
31	17	22	2,176	8,704	128	1,807,137,536
32	16	25	2,048	8,192	128	1,809,893,376
33	16	28	2,048	8,192	128	2,011,226,112
34	17	25	2,176	8,704	128	2,034,422,912
35	18	24	2,304	9,216	128	2,187,122,688
36	17	28	2,176	8,704	128	2,261,708,288
37	18	28	2,304	9,216	128	2,526,870,528
38	18	32	2,304	9,216	128	2,866,618,368
39	20	26	2,560	10,240	128	2,891,514,880
40	20	30	2,560	10,240	128	3,310,955,520
41	20	34	2,560	10,240	128	3,730,396,160
42	21	36	2,688	10,752	128	4,335,303,168
43	22	36	2,816	11,264	128	4,749,364,224
44	23	36	2,944	11,776	128	5,182,299,648
45	24	36	3,072	12,288	128	5,634,109,440
46	28	40	3,584	14,336	128	8,452,190,208
47	32	42	4,096	16,384	128	11,538,702,336
48	32	47	4,352	17,408	128	14,314,315,520

Continued on next page

Table C.3 continued from previous page

SCALE	HEADS	LAYERS	EMBED	FFW	KV	PARAMETERS
49	36	44	4,608	18,432	128	15,245,973,504
50	32	47	4,608	18,432	128	15,821,655,552
51	32	47	4,864	19,456	128	17,402,920,192
52	40	47	5,120	20,480	128	18,982,943,616

Table C.4: **The Unimax language sampling rates adapted from Chung et al. (2023).** Languages are listed in order of their percentage sampling rate, which sum to 100.

LANG	%	LANG	%	LANG	%	LANG	%	LANG	%	LANG	%
en	5.00e+00	af	1.42e+00	be	1.42e+00	fil	1.42e+00	gl	1.42e+00	te	1.42e+00
de	1.42e+00	es	1.42e+00	fr	1.42e+00	hr	1.42e+00	is	1.42e+00	kk	1.42e+00
ml	1.42e+00	mr	1.42e+00	ru	1.42e+00	sr	1.42e+00	ta	1.42e+00	az	1.42e+00
hi	1.42e+00	id	1.42e+00	it	1.42e+00	lv	1.42e+00	ms	1.42e+00	nl	1.42e+00
pl	1.42e+00	pt	1.42e+00	sq	1.42e+00	sv	1.42e+00	tr	1.42e+00	vi	1.42e+00
ca	1.42e+00	et	1.42e+00	hu	1.42e+00	iw	1.42e+00	ro	1.42e+00	sl	1.42e+00
th	1.42e+00	zh	1.42e+00	ar	1.42e+00	cs	1.42e+00	fa	1.42e+00	fi	1.42e+00
ja	1.42e+00	ko	1.42e+00	lt	1.42e+00	sk	1.42e+00	uk	1.42e+00	bg	1.42e+00
da	1.42e+00	el	1.42e+00	no	1.42e+00	mk	1.38e+00	bn	1.34e+00	eu	1.34e+00
ka	1.19e+00	mn	1.09e+00	bs	1.03e+00	uz	1.02e+00	ur	8.19e-01	sw	7.19e-01
ne	6.87e-01	kaa	6.76e-01	kn	6.72e-01	gu	6.43e-01	si	5.89e-01	cy	5.14e-01
eo	5.06e-01	la	4.64e-01	hy	4.47e-01	ky	4.37e-01	tg	4.28e-01	ga	4.25e-01
mt	4.06e-01	my	3.95e-01	km	3.35e-01	tt	3.14e-01	so	2.93e-01	ps	2.52e-01
ku	2.50e-01	pa	2.38e-01	rw	2.29e-01	lo	2.06e-01	dv	1.83e-01	ha	1.78e-01
ckb	1.73e-01	fy	1.68e-01	lb	1.63e-01	mg	1.54e-01	ug	1.52e-01	am	1.50e-01
gd	1.48e-01	ht	1.27e-01	grc	1.25e-01	jv	1.12e-01	tk	1.09e-01	hmn	1.09e-01
sd	1.05e-01	mi	9.77e-02	yi	9.55e-02	ba	9.42e-02	fo	9.24e-02	ceb	9.12e-02
or	9.07e-02	kl	8.12e-02	xh	7.21e-02	su	7.20e-02	ny	6.97e-02	sm	6.94e-02
sn	6.68e-02	co	6.67e-02	pap	6.57e-02	zu	6.46e-02	ig	6.31e-02	yo	6.00e-02
st	5.70e-02	haw	5.38e-02	as	5.07e-02	oc	4.93e-02	cv	4.66e-02	lus	4.61e-02
tet	4.14e-02	gsw	4.04e-02	sah	4.01e-02	br	3.29e-02	rm	2.52e-02	sa	2.23e-02
bo	2.23e-02	om	2.22e-02	se	2.12e-02	ce	1.70e-02	cnh	1.58e-02	ilo	1.49e-02
hil	1.44e-02	udm	1.39e-02	os	1.26e-02	lg	1.21e-02	ti	1.12e-02	vec	1.10e-02
ts	9.73e-03	tyv	9.66e-03	kbd	9.23e-03	ee	8.25e-03	iba	7.66e-03	av	7.57e-03
kha	7.57e-03	to	7.51e-03	tn	7.33e-03	nso	7.08e-03	fj	7.02e-03	zza	6.60e-03
ak	6.23e-03	ada	6.08e-03	otq	5.86e-03	dz	5.69e-03	bua	5.44e-03	cfm	5.41e-03

Continued on next page

Table C.4 continued from previous page

LANG	%	LANG	%	LANG	%	LANG	%	LANG	%	LANG	%
ln	5.39e-03	chm	5.36e-03	gn	5.23e-03	krc	5.21e-03	wa	5.11e-03	hif	4.79e-03
yua	4.32e-03	srn	4.26e-03	war	4.03e-03	rom	3.99e-03	bik	3.94e-03	sg	3.89e-03
lu	3.87e-03	ady	3.73e-03	kbp	3.68e-03	syr	3.51e-03	ltg	3.49e-03	myv	3.48e-03
iso	3.43e-03	kac	3.43e-03	bho	3.38e-03	ay	3.30e-03	kum	3.10e-03	qu	3.06e-03
pag	3.02e-03	ngu	2.97e-03	ve	2.94e-03	pck	2.88e-03	zap	2.86e-03	tyz	2.83e-03
hui	2.73e-03	bbc	2.65e-03	tzo	2.65e-03	tiv	2.55e-03	ksd	2.52e-03	gom	2.50e-03
min	2.47e-03	ang	2.46e-03	nhe	2.45e-03	bgp	2.45e-03	nzi	2.37e-03	nnb	2.29e-03
nv	2.28e-03	bci	2.26e-03	kv	2.25e-03	new	2.21e-03	mps	2.19e-03	alt	2.18e-03
meu	2.15e-03	bew	2.13e-03	fon	2.08e-03	iu	2.08e-03	abt	2.07e-03	mgh	2.05e-03
tvl	2.02e-03	dov	2.00e-03	tlh	1.96e-03	ho	1.96e-03	kw	1.92e-03	mrj	1.92e-03
meo	1.89e-03	crh	1.89e-03	mbt	1.87e-03	emp	1.85e-03	ace	1.85e-03	ium	1.85e-03
mam	1.81e-03	gym	1.74e-03	mai	1.72e-03	crs	1.70e-03	pon	1.69e-03	ubu	1.68e-03
quc	1.62e-03	gv	1.57e-03	kj	1.49e-03	btx	1.48e-03	ape	1.46e-03	chk	1.45e-03
rcf	1.44e-03	shn	1.42e-03	tzh	1.41e-03	mdf	1.39e-03	ppk	1.38e-03	ss	1.37e-03
gag	1.31e-03	cab	1.27e-03	kri	1.25e-03	seh	1.23e-03	ibb	1.23e-03	tbz	1.21e-03
bru	1.21e-03	enq	1.20e-03	ach	1.17e-03	cuk	1.16e-03	kmb	1.15e-03	wo	1.14e-03
kek	1.12e-03	qub	1.11e-03	tab	1.11e-03	bts	1.07e-03	kos	1.06e-03	rwo	1.05e-03
cak	1.05e-03	tuc	1.02e-03	bum	1.01e-03	gil	9.73e-04	stq	9.65e-04	tsg	9.47e-04
quh	9.39e-04	mak	9.37e-04	arn	9.35e-04	ban	9.06e-04	jiv	8.84e-04	sja	8.49e-04
yap	8.33e-04	tcy	8.26e-04	toj	8.20e-04	twu	8.15e-04	xal	8.08e-04	amu	8.05e-04
rmc	8.04e-04	hus	7.83e-04	nia	7.77e-04	kjh	7.73e-04	bm	7.64e-04	guh	7.64e-04
mas	7.64e-04	acf	7.56e-04	dtp	7.46e-04	ksw	7.25e-04	bzj	7.22e-04	din	7.17e-04
zne	7.15e-04	mad	6.99e-04	msi	6.84e-04	mag	6.60e-04	mkn	6.57e-04	kg	6.54e-04
lhu	6.37e-04	ch	6.23e-04	qvi	5.78e-04	mh	5.59e-04	djk	5.52e-04	sus	5.21e-04
mfe	5.20e-04	srm	5.12e-04	dyu	5.08e-04	ctu	5.07e-04	gui	5.03e-04	pau	5.02e-04
inb	4.88e-04	bi	4.71e-04	mni	4.59e-04	guc	4.43e-04	jam	4.39e-04	wal	4.38e-04
jac	4.35e-04	bas	4.30e-04	gor	4.18e-04	skr	4.15e-04	nyu	4.14e-04	noa	4.08e-04
sda	4.08e-04	gub	4.07e-04	nog	4.04e-04	teo	4.01e-04	tdx	3.92e-04	sxn	3.81e-04
rki	3.76e-04	nr	3.74e-04	frp	3.61e-04	alz	3.60e-04	taj	3.51e-04	lrc	3.50e-04
cce	3.22e-04	rn	3.21e-04	jvn	3.14e-04	hvn	3.08e-04	nij	3.07e-04	dwr	2.99e-04
izz	2.79e-04	msm	2.78e-04	bus	2.73e-04	ktu	2.66e-04	chr	2.52e-04	maz	2.39e-04
tzj	2.22e-04	suz	2.15e-04	knj	2.15e-04	bim	1.99e-04	gvl	1.98e-04	bqc	1.98e-04
tca	1.97e-04	pis	1.92e-04	laj	1.83e-04	qxr	1.82e-04	niq	1.80e-04	ahk	1.80e-04
shp	1.78e-04	hne	1.75e-04	spp	1.71e-04	koi	1.67e-04	quf	1.53e-04	agr	1.39e-04
tsc	1.31e-04	mqy	1.27e-04	gof	1.27e-04	gbm	1.25e-04	miq	1.22e-04	dje	1.21e-04
awa	1.19e-04	qvz	1.10e-04	tll	1.03e-04	raj	1.02e-04	kjg	1.00e-04	quy	9.24e-05
cbk	8.52e-05	akb	8.48e-05	oj	8.46e-05	ify	8.39e-05	cac	8.03e-05	brx	7.64e-05
qup	7.48e-05	ff	6.97e-05	ber	6.68e-05	tks	6.55e-05	trp	6.46e-05	mrw	6.46e-05

Continued on next page

Table C.4 continued from previous page

LANG	%	LANG	%	LANG	%	LANG	%	LANG	%	LANG	%
adh	6.40e-05	smt	6.16e-05	ffm	5.93e-05	qvc	5.87e-05	ann	5.40e-05	kaa_Latn	5.26e-05
nut	4.62e-05	kwi	4.34e-05	msb	4.20e-05	el_Latn	4.09e-05	doi	3.40e-05	dln	2.97e-05
hi_Latn	2.48e-05	ctd_Latn	1.03e-05	ru_Latn	6.95e-06	te_Latn	4.84e-06	ber_Latn	4.74e-06	az_RU	3.23e-06
ta_Latn	2.29e-06	tly_IR	2.15e-06	nan_Latn_TW	1.94e-06	ml_Latn	1.80e-06	zxx_xx_dtynoise	1.65e-06	gom_Latn	1.28e-06
bg_Latn	9.21e-07	kn_Latn	6.18e-07	zh_Latn	5.69e-07	cr_Latn	2.59e-07	bn_Latn	1.83e-07	gu_Latn	1.69e-07
sat_Latn	1.51e-07	ndc_ZW	1.31e-07	kmz_Latn	1.01e-07	ms_Arab	6.00e-08	ms_Arab_BN	4.29e-08		

C.2.3 Experimental Details: Scaling Laws Evaluation

Typically, to evaluate fitted scaling laws, prior work has adopted the R^2 metric, for it's scale-invariant measure of a laws' explanatory power over the observed data. However, the choice of what data to hold-out as the test set can have a significant impact on the performance and it's interpretation (Li et al., 2025a). For this reason, and because multilingual scaling laws are used to measure more than just the predictive power of the law to larger scales, we use a set of different R^2 metrics to measure different objectives. These are:

1. R^2 Hold-out a randomly selected 20% of the data points from the data. This gives an impression of the models' ability to fit the variance, without measuring particular dimensions of fit.

2. $R^2(D)$ Hold-out the training runs with training tokens with the top 20% of D tokens.

3. $R^2(N)$ — Hold-out a couple of scales of model sizes, including the largest: $N = 660M$ and $N = 8B$ parameters.

4. $R^2(C)$ Hold-out the largest compute parts of the training runs, where $C = 6ND$ FLOPs. This will include mainly a combination of the larger model sizes, and longer training runs.

5. $R^2(M)$ Hold-out all mixtures that are not monolingual, bilingual, or unimax. This allows the scaling law to get a baseline sense of each interacting language, but it has to infer the rest of the mixture scaling properties.

These metric variants allow us to unpack specifically where the scaling laws perform well, and for what purposes they are reliable.

C.2.4 Experimental Details: Language Finetuning

In Subsection 4.1.5 we experiment with continuous pretraining on a single language, starting from a massively multilingual model's checkpoint. For clarity, we refer to this as *finetuning*, to distinguish it from pretraining from scratch.

We begin by training massively multilingual models, using UNIMAX (Chung et al., 2023) language sampling across all 420 languages in MADLAD-400. Chung et al. (2023) demonstrate this is an effective sampling strategy to perform well across languages, as is the objective with many large multilingual models. The UNIMAX sampling rate per language is documented in Table C.4 Each of these models we train for $1T$ tokens, to reflect real-world practices.

Using these Unimax checkpoints, across model scales, we then conduct continuous monolingual pretraining (which we call *finetuning* here) on all 48 languages separately in $\mathcal{L}_{\text{eval}}$ (48). While finetuning, we also observe the loss across all $\mathcal{L}_{\text{eval}}$ (48)languages as well. We use the same hyperparameters as discussed for pretraining, but reset the learning schedule, and train for at least another $30k+$ steps. These empirical results allow us to compare (across a range of N, D, C) the loss on language t between a model finetuned from a Unimax checkpoint, and a model pretrained monolingually from scratch on language t.

C.2.5 Experimental Details: Language Transfer

In Subsection 4.1.3 we measure how much training on *source language* s is beneficial for the test loss on the *target language* t. A language transfer score tells a practitioner the extent to which training with each language would help or hinder their target language t. There are two ways in which we can measure this language transfer:

1. BILINGUAL TRANSFER SCORE. This score measures how many training tokens a 50/50 (s, t) bilingual model needs, relative to a monolingual target language t baseline, to reach the same validation loss L_t. A positive score indicates co-training with the source language has positive transfer, whereas a negative score indicates it hinders convergence.
2. FINETUNING ADAPTATION SCORE. This score captures the effect on the target language's loss L_t, when a massively multilingual model is finetuned only on the source language s.

C.2.5.1 Bilingual Transfer Score (BTS)

The Bilingual Transfer Score asks how data-efficient a bilingual learner is, relative to a purely monolingual one, on some target language t. To estimate this, we pretrain two models, a monolingual one on t and a bilingual one that samples tokens equally (50/50) from the source and target languages (s, t). We then measure how many more training tokens it takes the bilingual model to attain the same validation loss L_t as the monolingual model. This "distance" between the learning curves is measured throughout training, at different reference horizons—or, number of tokens trained on. In Figure 4.2 we use models with 2B parameters, and a reference horizon of $d = 42B$ tokens, as it roughly equates to $10,000$ pretraining steps for our 2B parameter models, and learning curves have stabilized. In Figure C.1 we measure how the BTS varies by model size, and by number of training tokens seen.

We are able to compute the bilingual transfer scores, across every combination of pairings between 10 languages: English, Russian, Spanish, French, German, Portuguese, Chinese, Vietnamese, Japanese, and Hindi. These languages were chosen to give a distribution of high- and mid-resource languages from a variety of language families. This totals 10 monolingual experiments (one for each language), and $C(10, 2) = 45$ bilingual experiments, which provide the results for a 10x10 grid of Bilingual Transfer Scores.

In more detail, let d_{mono} be the number of tokens trained on language t by a monolingual model. This model attains test loss of $L_t(d_{mono})$ after d_{mono} tokens on language t. We record the number of training tokens d_{bi} at which the bilingual model reaches the same loss, $L_t(d_{\text{bi}}) = L_t(d_{mono})$. We define the **Bilingual Transfer Score** (BTS) as the signed step surplus

$$\text{BTS}_{s \to t} = -\frac{\sigma_{\text{bi}}(L_{\text{t}}(d_{\text{mono}})) - 2d_{\text{mono}}}{d_{\text{mono}}}$$

where $d_{\text{bi}} = \sigma_{\text{bi}}(L_{\text{t}}(d_{\text{mono}}))$. σ_{bi} maps a loss value to the number of steps taken by the bilingual model to reach that loss. In short, this score measures the relative training efficiency of the bilingual model compared to a monolingual model at reaching the same loss level for language t. Subtracting $2d_{\text{mono}}$ and dividing by d_{mono} simply centers the value on 0, and forces positive transfer to give a positive value. A value of 0, implies *neutral* transfer, as the multilingual model requires exactly the same number of steps in the target language t as the monolingual model, a value below 0 implies *negative* transfer, as the multilingual model is slower to reach the loss $L_{\text{t}}(d_{\text{mono}})$, and a value above 0 implies *positive* transfer, as the bilingual model reaches the target loss in $< 2d_{\text{mono}}$. Note that for Figure 4.2 these scores are anchored on 2B parameter models, trained

for $d = 42B$ tokens, for consistency. But as model size increases, or tokens trained decreases, the language transfer scores improve monotonically.

C.2.5.2 Finetuning Adaptation Score (FAS)

Complementary to BTS, the Finetuning Adaptation Score measures the *instantaneous impact* of continued exposure to a single language on all others while holding model capacity fixed. Specifically, it measures how finetuning a massively multilingual model on source language s affects the performance of target language t. Because we only need to train on each source language once, and evaluate on all languages, it is less computationally expensive than BTS, which requires training on every pair of languages. As such, we are able to derive $38 * 38 = 1444$ (s, t) comparisons from 38 finetuning experiments.

We start from a shared multilingual checkpoint: a Unimax pretrained model, trained for $1T$ tokens. For each source language s, we continue pre-training (for simplicity we denote this as *finetune*) exclusively on this language, and evaluate at regular intervals on every target language's $t \in \mathcal{L}$ validation set. Let $L_{s \to t}(d)$ be the validation loss on t after d additional updates on s, and let L_t^{unimax} denote the baseline loss of the shared Unimax checkpoint, before finetuning has begun. The Finetuning Adaptation Score (FAS) aggregates the reduction in loss over a common time window $[0, d_{\max}]$:

$$\mathrm{FAS}_{s \to t} = \frac{1}{d_{\max}} \int_0^{d_{\max}} \left[L_t^{unimax} - L_{s \to t}(d) \right] \mathrm{d}d.$$

Positive values indicate that exposing the model to s accelerates learning or yields immediate gains on t ("helpful transfer"), whereas negative values capture negative interference. Normalising by $d_{\max}$ makes the score comparable across language pairs and training horizons.

Because finetuning on language s can cause unpredictable behavior in the validation loss of a target language t (it might immediately increase, or decrease then eventually increase), we start with deriving a series of features about the finetuning loss curve. We calculate the following features for a finetuning loss curve:

- **Baseline Loss Deviation ($\delta_{s \to t}$).** For each language pair $s \neq t$ we quantify the deviation in validation loss between the baseline L_t^{unimax} and the loss $L_{s \to t}$ on the final step:

  $$\delta_{s \to t} = L_{s \to t}(d_{\max}) - L_t^{unimax}(d_{\max})$$

  $\delta_{s \to t} < 0$ means fine–tuning on s helps t; $\delta_{s \to t} > 0$ means it hurts.

- **Adaptation Gain (g_l).** This measures the area normalized reduction in loss on language l from finetuning on language l. This allows us to capture the relative learning dynamics of both the source and target language.

$$g_l = \frac{1}{d_{max}} \int_0^{d_{\max}} \left[L_l^{unimax} - L_{s \to t}(d) \right] \, \mathrm{d}s.$$

  $g_l > 0$ indicates net improvement; large values imply that l benefits greatly from additional supervision.

For each language transfer pair (s, t), we compose the following feature vector $x_{s \to t}$. As languages differ in intrinsic difficulty, we normalize each feature.

$$\tilde{g}_l = \frac{g_l - \mu_g}{\sigma_g}, \qquad \tilde{\delta}_{s \to t} = \frac{\delta_{s \to t} - \mu_{\delta,t}}{\sigma_{\delta,t}}, \tag{C.2}$$

For adaptation gain, we use the global mean and standard deviation (μ_g, σ_g), and for baseline loss deviation we compute the mean and standard deviation $(\mu_{\delta,t}, \sigma_{\delta,t})$ across all source languages s for each fixed target language t.

C.2.6 Estimating Bilingual Transfer Score (BTS)

In the Bilingual Transfer Score experiments we empirically measured language transfer scores $y_{s \to t}^{\text{true}}$ for a subset $\mathcal{P}_{\text{obs}} \subset \mathcal{L} \times \mathcal{L}$. We construct a training set from these 90 experiments, using them as the gold labels.

$$\mathbf{x}_{s \to t} = \left[\tilde{g}_s, \ \tilde{g}_t, \ \tilde{g}_s - \tilde{g}_t, \ \tilde{\delta}_{s \to t} \right]^\top, \qquad y_{s \to t} = y_{s \to t}^{\text{true}}, \tag{C.3}$$

We fit a random-forest regressor $\varphi_{\boldsymbol{\theta}} : \mathbb{R}^4 \to \mathbb{R}$ with 300 trees and unlimited depth. The predictive performance is assessed by k-fold cross-validation ($k = 5$), obtaining an $R^2 = 0.85$ and Spearman correlation of $\rho = 0.88$. After training on all observed pairs, φ is invoked on every unlabelled $(s, t) \in \mathcal{L} \times \mathcal{L}$, yielding predicted BTS values, that estimate the language transfer shown in Figure 4.2.

C.2.7 Experimental Details: Multilingual Capacity

C.2.7.1 Multilingual Capacity Scaling Laws

In Subsection 4.1.4 we aim to understand how a model's capacity hinders multilingual learning—also termed the *curse of multilinguality* (Conneau et al., 2019; Chang et al., 2024). To do this, we train models of various sizes, and with a range of language mixtures, shown in Table C.5. For simplicity, languages are always evenly

sampled within their mixtures, even if some languages have more available text than others. We train models on these mixtures, and evaluate the validation loss of all other languages, throughout training.

For a given target language t we have empirical loss scores across 11 differently sized models (from 10M to 8B), and across each mixture that contains the language in training, including monolingual, bilingual, and the multilingual mixtures shown in Table C.5.

We model each language's loss as $L(N, D, K)$, where N is the model size, D are the number of training tokens for the *target* language, and K is the number of languages in the training mixture (evenly sampled). We define $L(N, D, K)$ as follows:

$$L(N, D, K) \;=\; L_\infty \;+\; \frac{A\,K^\phi}{N^\alpha} \;+\; \frac{B\,K^\psi}{D^\beta} \tag{C.4}$$

This law preserves the well-validated power-law behavior of monolingual scaling laws, while introducing K into both the capacity term (AK^ϕ/N^α) and the data term (BK^ψ/D^β). The exponent ϕ captures how capacity requirements grow with the number of languages, while ψ captures how data requirements change with multilinguality: $\psi < 0$ indicates *positive transfer* (sublinear per-language data needs as K increases), $\psi = 0$ implies no net transfer in the data term, and $\psi > 0$ implies negative transfer/interference. Under even sampling, the total tokens across languages are $D_{\text{tot}} = K \cdot D$, in which case the data term can be rewritten as $B\,K^{\psi+\beta}/D_{\text{tot}}^\beta$.

C.2.7.2 Adding languages: K to $K * r$

Practitioners may wish to change the number of languages supported by a model, from K to $K * r$. If the model size and training tokens remain unchanged, it will lead to a degradation in loss across languages, as the same compute budget is allocated across more languages. When adding languages, the practitioner may wish to know how much they need to scale the model (N, D) and the training compute C to maintain the same performance per language, as before. We call this an *iso-loss*, as it approximately preserves the loss values, with a change in K.

Compute-Optimality for Equation (C.4). As we increase K to $K * r$ we would like to understand the new N', D' that *minimize* training compute on the iso-loss surface. Let $r = K'/K$. Minimizing $C' \propto N'D'$ subject to

$$\frac{A(Kr)^\phi}{N'^\alpha} + \frac{B(Kr)^\psi}{D'^\beta} \;=\; \frac{AK^\phi}{N^\alpha} + \frac{BK^\psi}{D^\beta}$$

yields the closed-form optimum:

$$\left(\frac{N'}{N}\right)^\star = r^{\phi/\alpha}, \qquad \left(\frac{D'}{D}\right)^\star = r^{\psi/\beta}, \qquad \left(\frac{D'_{tot}}{D_{tot}}\right)^\star = r^{1+\psi/\beta}.$$

These expressions show a clear separation of roles: ϕ/α governs the parameter burden of adding languages, whereas ψ/β governs the data burden (with $\psi < 0$ reducing the per-language data requirement due to positive transfer). Since the compute budget is in terms of total tokens $D_{\text{tot}} = K \cdot D$, and $C \propto N D_{\text{tot}}$.

Adding languages: K to $K * r$ — Finding the compute multiplier. At the iso-loss optimum, the compute multiplier is

$$\frac{C'}{C} \;=\; \frac{N' D'_{tot}}{N D_{tot}} \;=\; \left(\frac{N'}{N}\right)^\star \left(\frac{D'_{\text{tot}}}{D_{\text{tot}}}\right)^\star = r^{1+\phi/\alpha+\psi/\beta}\,.$$

Intuitively, $\frac{\phi}{\alpha}$ measures the extra parameter capacity per added language, while $\frac{\psi}{\beta}$ measures the extra (or reduced, if $\psi < 0$) per-language token budget. But since compute is defined with total tokens D_{tot}, then we consider the additional factor of r from enlarging the language inventory.

Adding languages: K to $K * r$ — Finding the Iso-loss curve. Let $s \equiv \frac{N'}{N}$ and $t \equiv \frac{D'}{D}$ denote the multipliers for model size and (per-target-language) data when we grow the language inventory from K to $K' = rK$. Define the baseline term contributions

$$w_N \;\equiv\; \frac{AK^\phi/N^\alpha}{AK^\phi/N^\alpha + BK^\psi/D^\beta}, \qquad w_D \;=\; 1 - w_N.$$

The iso-loss condition equates the *sum of the two error terms* before and after the change in K:

$$\frac{A(Kr)^\phi}{(Ns)^\alpha} \;+\; \frac{B(Kr)^\psi}{(Dt)^\beta} \;=\; \frac{AK^\phi}{N^\alpha} \;+\; \frac{BK^\psi}{D^\beta}.$$

Dividing both sides by $AK^\phi/N^\alpha + BK^\psi/D^\beta$ yields the normalized *iso-loss constraint*

$$r^\phi\, w_N\, s^{-\alpha} \;+\; r^\psi\, w_D\, t^{-\beta} \;=\; 1. \tag{C.5}$$

Solving for t given s. Rearranging equation C.5 gives

$$t \;=\; \left[\frac{r^\psi\, w_D}{1 - r^\phi\, w_N\, s^{-\alpha}}\right]^{1/\beta}.$$

The denominator must be positive, which implies the feasibility condition

$$s^\alpha \;>\; r^\phi\, w_N.$$

As $s^\alpha \downarrow r^\phi w_N$ the denominator tends to 0^+ and $t \to \infty$; as $s \to \infty$, the denominator tends to 1 and $t \to (r^\psi w_D)^{1/\beta}$.

Solving for s given t. By symmetry,

$$s \;=\; \left[\frac{r^\phi\, w_N}{1 - r^\psi\, w_D\, t^{-\beta}}\right]^{1/\alpha}, \qquad \text{with feasibility } \; t^\beta \;>\; r^\psi\, w_D.$$

Total-token form. Under even sampling, $D_{\text{tot}} = KD$ and $D'_{\text{tot}} = K'D' = rKD'$, so

$$\frac{D'_{\text{tot}}}{D_{\text{tot}}} \;=\; r\,\frac{D'}{D} \;=\; r\left[\frac{r^\psi\, w_D}{1 - r^\phi\, w_N\, s^{-\alpha}}\right]^{1/\beta}.$$

Compute-optimal frontier (starting from a compute-optimal (K, N, D)). Suppose the *baseline* (K, N, D) is compute-optimal for its loss level, i.e., it minimizes $C \propto ND$ subject to $AK^\phi/N^\alpha + BK^\psi/D^\beta = \text{const}$. A standard Lagrange multiplier argument then yields

$$\alpha\,\frac{AK^\phi}{N^\alpha} \;=\; \beta\,\frac{BK^\psi}{D^\beta} \quad\Longleftrightarrow\quad w_N \;=\; \frac{\beta}{\alpha+\beta}, \quad w_D \;=\; \frac{\alpha}{\alpha+\beta}.$$

After changing $K \to rK$, the compute along the iso-loss curve is $C'/C = s\,t$ (or $C'/C = s \cdot \frac{D'_{\text{tot}}}{D_{\text{tot}}}$ if compute is defined with total tokens). Minimizing this product subject to equation C.5 again via Lagrange multipliers gives the stationarity condition

$$\alpha\, r^\phi\, w_N\, s^{-\alpha} \;=\; \beta\, r^\psi\, w_D\, t^{-\beta}.$$

Combining with equation C.5 and substituting the compute-optimal weights above yields the *unique* minimum-compute point on the new frontier:

$$\left(\tfrac{N'}{N}\right)^\star = r^{\phi/\alpha}, \qquad \left(\tfrac{D'}{D}\right)^\star = r^{\psi/\beta}, \qquad \left(\tfrac{D'_{\text{tot}}}{D_{\text{tot}}}\right)^\star = r^{\,1+\psi/\beta}.$$

At this point each error component is *individually restored* to its baseline magnitude: $A(Kr)^\phi/N'^\alpha = AK^\phi/N^\alpha$ and $B(Kr)^\psi/D'^\beta = BK^\psi/D^\beta$, so the weights (w_N, w_D)—and hence the balance of capacity/data bottlenecks—remain unchanged. Because the frontier is convex in $(\log s, \log t)$ and $\log C' = \log s + \log t$ is linear, this stationary point is the global compute minimum along the iso-loss curve.

C.3 Extended Results

C.3.1 Extended Results: Scaling Laws

In Table C.6 we illustrate the fitted scaling laws for each language. Specifically, we show the scaling law parameters for each language when trained with a monolingual vocabulary on monolingual data, when trained

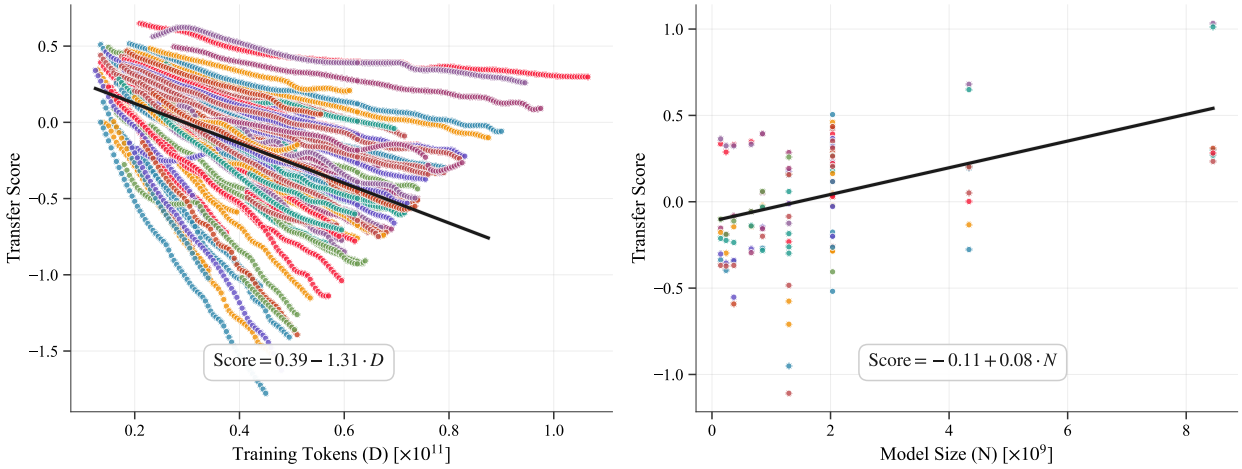


Figure C.1: **Left:** X-axis is training tokens (D), Y-axis is language-transfer score. **Right:** X-axis is model size (N), Y-axis is language-transfer score.

with a multilingual vocabulary on monolingual data, and when trained with a multilingual vocabulary on the Unimax multilingual mixture. These results match up with those presented in Figure 4.1.

C.3.2 Extended Results: Language Transfer

Research Question: *How does language transfer change with larger models, or when you train for longer?*

We further investigated how language transfer evolves with training data (D) and model size (N) in Figure C.1. Our analysis of transfer scores over the course of training reveals that transfer dynamics are established early. The performance gap between synergistic pairs (e.g., es $\rightarrow$ pt) and interfering pairs (e.g., en $\rightarrow$ zh) appears within the initial stages of training and remains largely consistent, suggesting that the fundamental compatibility of languages is not significantly altered by longer training. However, the impact of model scale is more pronounced. As illustrated in Figure C.1, we observe a clear trend in which larger models are more effective in facilitating cross-lingual transfer. For synergistic pairs, the positive transfer score increases modestly with model capacity. More importantly, for challenging pairs that exhibit interference in smaller models, larger models are significantly better at mitigating this negative effect, bringing the score closer to zero. This is similar to what has been observed while training larger and larger foundation models (Team et al., 2024). Given that transfer dynamics change at scale, it is an important factor to consider when designing multilingual foundation models that are performant for all languages being trained on.

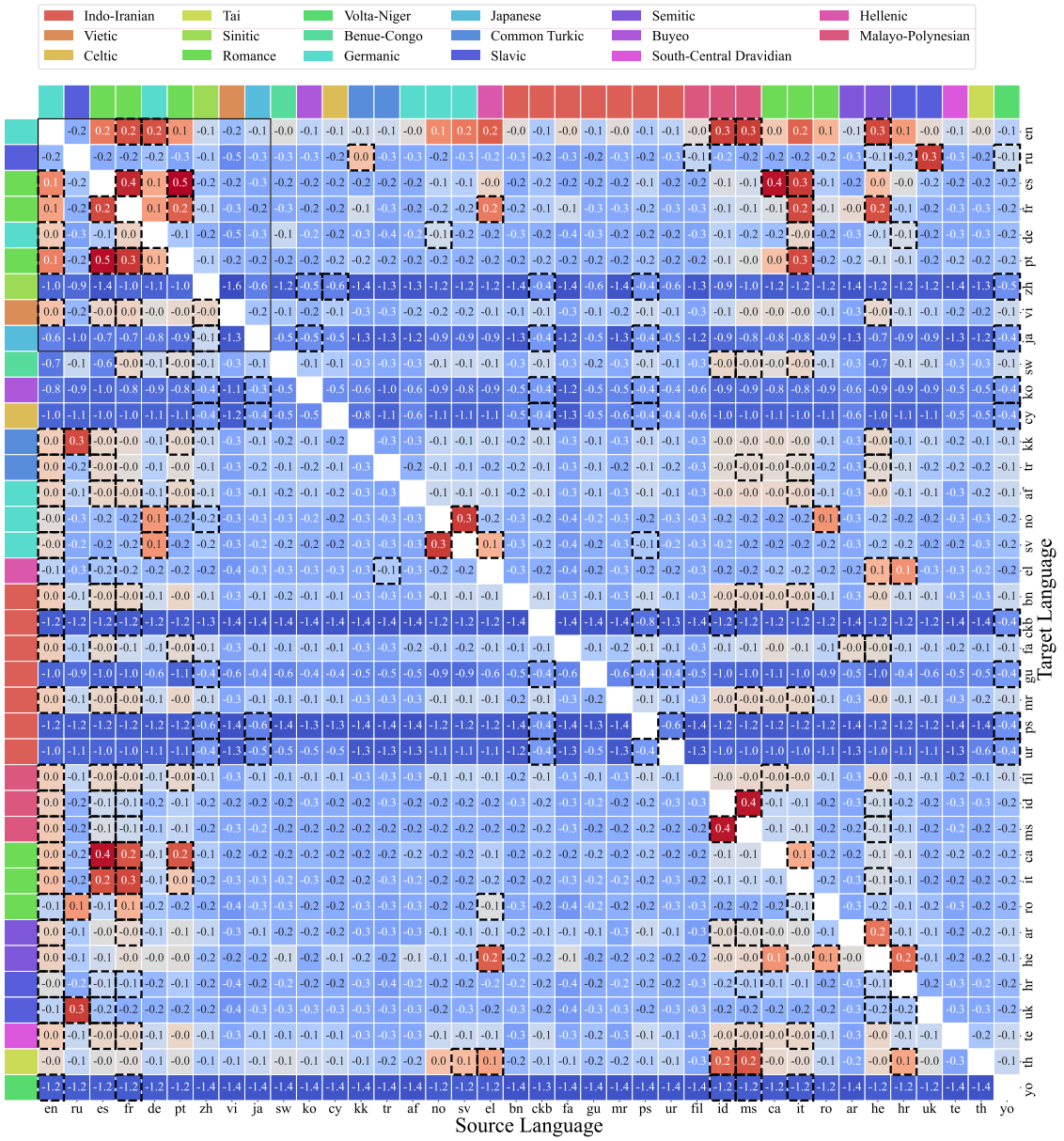


Figure C.2: The language transfer scores measure the benefit to the target language of (co-)training with the source language. We bold the top-5 source languages for each target language. The top left 9x9 are the Bilingual Transfer Score computed directly, whereas the rest are estimated from the Finetuning Adaptation Score. While English is the best source language for many of of languages, we find language similarity is highly predictive of these scores.

Table C.1: **An overview of experiment configurations in this work**. We enumerate the experiment types: **<Lang>** for monolingual scaling, **Unimax** as a massively multilingual baseline, **Language Pairs** to measure language-to-language transfer, **Capacity** to measure the curse of multilinguality, or model capacity constraints on learning new languages, and **Finetunes** to understand how finetuning from a massively multilingual model compares to pretraining from scratch. We use a mix of Monolingual and Multilingual vocabularies, and training data. In the LANGUAGES and SCALES columns we use parentheses to show the number of language mixtures and number of scales run. Symbols such as $\mathcal{L}_{\text{mono}}$ (7) and $\mathcal{S}_{\text{full}}$ (20) are defined above.

EXPERIMENT TAG	VOCAB.	TRAIN-DATA	LANGUAGES	SCALES	# EXPS
Mono <Lang>	Monolingual	Monolingual	$\mathcal{L}_{\text{mono}}$ (7)	$\mathcal{S}_{\text{full}}$ (20)	$7 \times 20 = 140$
Unimax	Multilingual	Multilingual	$\mathcal{L}_{\text{unimax}}$	$\mathcal{S}_{\text{full}}$ (20)	20
Multi <Lang>	Multilingual	Monolingual	$\mathcal{L}_{\text{mono}}$ (7)	$\mathcal{S}_{\text{part}}$ (11)	$10 \times 7 = 70$
Language Pairs	Multilingual	Monolingual	$\mathcal{L}_{\text{eval}}$ (48)	34	50
Language Pairs	Multilingual	Bilingual	$\mathcal{L}_{\text{pairs}}$ (10)	34	90
Language Pairs	Multilingual	Bilingual	$\mathcal{L}_{\text{target}}$ (15)	$\mathcal{S}_{\text{part}}$ (11)	$10 \times 15 = 150$
Capacity	Multilingual	Uniform Sampling	$\mathcal{C}_{\text{exp}}$ (12)	$\mathcal{S}_{\text{part}}$ (11)	$10 \times 12 = 120$
Finetunes	Multilingual	Monolingual	$\mathcal{L}_{\text{eval}}$ (48)	34	50
Finetunes	Multilingual	Monolingual	$\mathcal{L}_{\text{mono}}$ (7)$\cup\mathcal{L}_{\text{pairs}}$ (10)	$\mathcal{S}_{\text{min}}$ (7)	$7 \times 12 = 84$
				Total	774

Table C.5: **The set of experiments designed to measure the *curse of multilinguality*—or how language performance is impacted by both the number of training languages and the model size.** We defined VERSION mixtures with NUM languages, ranging between 4 and 50. For mixtures with 4 languages, we define a few different versions oriented to languages that might be more likely to be grouped together. This set of experiments are summarized in the **Capacity** row of Table C.1.

VERSION	NUM	LANGUAGES (ALPHABETICAL)
4v0	4	de, es, fr, pt
4v1	4	de, en, es, fr
4v2	4	hi, ja, vi, zh
4v3	4	en, hi, pt, ru
6v0	6	de, en, es, fr, pt, ru
8v0	8	de, en, es, fr, pt, ru, vi, zh
8v1	8	de, en, fr, hi, ja, ru, vi, zh
12v0	12	de, en, es, fr, hi, it, ja, pl, pt, ru, vi, zh
16v0	16	de, en, es, fr, hi, id, it, ja, nl, pl, pt, ru, sv, tr, vi, zh
24v0	24	cs, da, de, en, es, fa, fi, fr, hi, hu, id, it, ja, nl, pl, pt, ro, ru, sv, th, tr, uk, vi, zh
32v0	32	ar, bg, ca, cs, da, de, el, en, es, fa, fi, fr, hi, hu, id, it, ja, ko, lt, nl, no, pl, pt, ro, ru, sk, sv, th, tr, uk, vi, zh
50v0	50	af, ar, ba, bm, bn, ca, ckb, cy, de, ee, el, en, es, fa, fil, fr, gn, gu, ha, he, hi, hr, id, it, ja, jv, kk, ko, mr, ms, my, nl, no, om, pl, ps, pt, ro, ru, sm, sv, sw, te, th, tr, uk, ur, vi, yo, zh

Table C.6: **A summary of the monolingual scaling laws for each language** The columns are: the language, the language family, the language script, its number of unique tokens $|D|$ in MADLAD-400 (with the percentage of unique tokens as compared to English), and the derived scaling law.

LANGUAGE	FAMILY	SCRIPT	$\lvert D\rvert$ (% EN)	SCALING LAW
Monolingual Vocabulary, Monolingual Training				
English	Indo-european	Latin	2.8T (100.0%)	$0.67 + \frac{e^{6.18}}{N^{0.41}} + \frac{e^{8.25}}{D^{0.41}}$
French	Indo-european	Latin	363B (13.01%)	$0.76 + \frac{e^{10.11}}{N^{0.64}} + \frac{e^{13.24}}{D^{0.64}}$
Russian	Indo-european	Cyrillic	705B (25.26%)	$0.82 + \frac{e^{10.07}}{N^{0.63}} + \frac{e^{13.28}}{D^{0.63}}$
Chinese	Sino-tibetan	Hans	125B (4.48%)	$1.08 + \frac{e^{8.20}}{N^{0.46}} + \frac{e^{10.37}}{D^{0.46}}$
Hindi	Indo-european	Devanagari	7.9B (0.28%)	$0.52 + \frac{e^{6.03}}{N^{0.39}} + \frac{e^{8.03}}{D^{0.39}}$
Swahili	Niger-congo: atlantic congo	Latin	770M (0.03%)	$0.00 + \frac{e^{4.13}}{N^{0.26}} + \frac{e^{5.51}}{D^{0.26}}$
Multilingual Vocabulary, Monolingual Training				
English	Indo-european	Latin	2.8T (100.0%)	$0.83 + \frac{e^{9.91}}{N^{0.63}} + \frac{e^{13.06}}{D^{0.63}}$
French	Indo-european	Latin	363B (13.01%)	$0.66 + \frac{e^{7.12}}{N^{0.45}} + \frac{e^{8.94}}{D^{0.45}}$
Russian	Indo-european	Cyrillic	705B (25.26%)	$0.76 + \frac{e^{7.97}}{N^{0.49}} + \frac{e^{9.91}}{D^{0.49}}$
Chinese	Sino-tibetan	Hans	125B (4.48%)	$1.18 + \frac{e^{8.87}}{N^{0.49}} + \frac{e^{10.90}}{D^{0.49}}$
Hindi	Indo-european	Devanagari	7.9B (0.28%)	$0.63 + \frac{e^{6.99}}{N^{0.43}} + \frac{e^{8.69}}{D^{0.43}}$
Swahili	Niger-congo: atlantic congo	Latin	770M (0.03%)	$0.00 + \frac{e^{4.99}}{N^{0.30}} + \frac{e^{6.43}}{D^{0.30}}$
Multilingual Vocabulary, Unimax Training				
English	Indo-european	Latin	2.8T (100.0%)	$0.00 + \frac{e^{3.03}}{N^{0.16}} + \frac{e^{3.15}}{D^{0.16}}$
French	Indo-european	Latin	363B (13.01%)	$0.00 + \frac{e^{3.63}}{N^{0.19}} + \frac{e^{3.94}}{D^{0.19}}$
Russian	Indo-european	Cyrillic	705B (25.26%)	$0.00 + \frac{e^{3.90}}{N^{0.20}} + \frac{e^{4.28}}{D^{0.20}}$
Chinese	Sino-tibetan	Hans	125B (4.48%)	$0.39 + \frac{e^{4.58}}{N^{0.21}} + \frac{e^{5.47}}{D^{0.21}}$
Hindi	Indo-european	Devanagari	7.9B (0.28%)	$0.00 + \frac{e^{3.46}}{N^{0.18}} + \frac{e^{3.89}}{D^{0.18}}$
Swahili	Niger-congo: atlantic congo	Latin	770M (0.03%)	$0.00 + \frac{e^{3.52}}{N^{0.18}} + \frac{e^{3.74}}{D^{0.18}}$