Training Bilingual LMs with Data Constraints in the Targeted Language

Skyler Seto[*], **Maartje ter Hoeve**[*], **Richard He Bai, Natalie Schluter, David Grangier**
Apple
{sseto,m_terhoeve,richardbai,natschluter,d_grangier}@apple.com

Abstract

Large language models are trained on massive scrapes of the web, as required by current scaling laws. Most progress is made for English, given its abundance of high-quality pretraining data. For most other languages, however, such high quality pretraining data is unavailable. In this work, we study how to boost pretrained model performance in a target language with insufficient pretraining data for training a high performing language model, by enlisting data from an auxiliary language for which high quality data is available. We study this by quantifying the performance gap between training with data in a data-rich auxiliary language compared with training in the target language, exploring the benefits of translation systems, studying the limitations of model scaling when data is limited in the target languages, and proposing new methods for upsampling data from the auxiliary language. Our results show that stronger auxiliary datasets result in performance gains without modification to the model or training objective for close languages.

1 Introduction

The abundance of high quality English data has given rise to exceptional language modeling performance in English (Brown et al., 2020; Bubeck et al., 2023; OpenAI, 2023). However, there are many other languages for which a reasonably large amount of data is available, yet not in equally large amounts as for English. Taking mC4 (Xue, 2020) as an example, languages like German, French, Chinese, and Japanese have ∼10-100 times less data than English. Consequently, most non-English progress comes from relatively small bilingual models (e.g., Le et al., 2019; De Vries et al., 2019; Martin et al., 2019; Scheible et al., 2020; Wei et al., 2023a; Faysse et al., 2024), or larger massively multilingual models (e.g., Le Scao et al., 2023; Intrator

et al., 2024; Üstün et al., 2024). Other LLMs such as Llama-2, GPT-3, and PaLM-2 that perform well across a variety of languages are trained primarily on English, with less than 20% of data from other languages (Xu et al., 2024).

Several works study cross-lingual transfer, but are limited to finetuning smaller models (∼100M parameters) with limited data, or classical NLP tasks, such as part-of-speech tagging, named entity recognition, and natural language inference (e.g., Chang et al., 2023; de Vries et al., 2022; Faisal and Anastasopoulos, 2024). They do not evaluate pretrained language models on contemporary knowledge-based downstream tasks. It is still unclear how to optimally make use of bi- or multilingual data, to optimally increase performance in a target language. This is especially relevant in data constrained settings. For example, imagine one has access to only 500K documents in a target language, and wants to pretrain a language model for this language. 500K documents is sufficient to attain a reasonable perplexity score, but this amount of data is insufficient to do well on modern knowledge-based downstream tasks. Many languages in the tail of mC4 fall in this category. Currently, a systematic analysis of how to use an auxiliary language in this scenario is lacking, leaving practitioners to a trial-and-error based approach, mostly based on intuition.

In this work, we provide such a systematic analysis. In particular, we are interested in whether "higher quality" auxiliary data also leads to higher model performance in the target language. We choose English as the main auxiliary language, matching a common practical setting. To investigate the effect of linguistic similarity, we also experiment with Chinese as the auxiliary language, although being somewhat bottlenecked by the amount of available data in Chinese.

To measure the impact of "better" auxiliary data, we explore well-tested filtering techniques for a
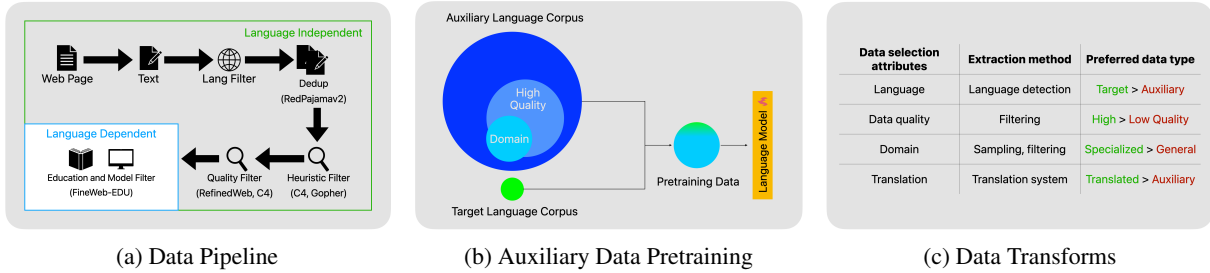
---

[*]Equal contribution

Figure 1: (a) **Data Pipeline**: English data pipeline used for building large pretraining corpora in (Penedo et al., 2024). (b) **Auxiliary Data Pretraining**: Combining high quality domain-specific pretraining data with a small amount of data from the target language for pretraining with limited target data. (c) **Data Transforms**: Many considerations when building datasets in languages with limited data.

monolingual, English, setting. We focus on the impact of dataset size, filtering for data quality and style, and data selection for specialized information relevant to downstream evaluations (Figure 1c), matching recent advancements in state-of-the-art open-source English datasets such as FineWebEDU (Penedo et al., 2024), and DataComp (Li et al., 2024) (Figure 1). Currently, it is still unclear how these techniques extend to a bilingual setting; as we confirm in our experiments, not every filtering technique that performs well in a monolingual setting performs well in a bilingual setting. Specifically, our **key findings** are:

1. Auxiliary English data that is generated by some of the existing model-based data filtering pipelines for English *can* be helpful to complement limited data in a target language (§3.2);

2. Findings are not the same across multiple target languages. We hypothesize that for languages that are "far" from English, gains from better English datasets do not help (§3.3);

3. When training with high quality auxiliary data, there is a marginal gap between upsampling *relevant information* and filtering for high quality instructional text (§4.1-4.2);

4. There are limits to the model sizes that are practical to pretrain with limited target data: data size should scale linearly with model size (Kaplan et al., 2020), and performance in the target language saturates without increasing target data, regardless of increasing auxiliary data (§5.2).

2 Related Work

Multilingual Language Models While much of the LLM research has focused on English, large-scale transformer-based multilingual language models have been trained on large multilingual corpora including mBERT (Pires, 2019),

XLM (Conneau and Lample, 2019), mT5 (Xue, 2020), PolyLM (Wei et al., 2023b), and Bloom (Le Scao et al., 2023). These works focus on training models that are language balanced and that perform well across multilingual benchmark tasks in over 100 languages. Other SOTA language models such as Llama 2 (Touvron et al., 2023), Falcon (Almazrouei et al., 2023), and Palm 2 (Anil et al., 2023) have multilingual capabilities, but over 90% the training data is English, and these models perform poorly across a variety of languages, such as south east Asian languages (Nguyen et al., 2023).

Other works focus on training smaller bilingual language models in French (Faysse et al., 2024; Le et al., 2019; Martin et al., 2019), German (Scheible et al., 2020), Dutch (De Vries et al., 2019), or Chinese (Wei et al., 2023a), but require a substantial amount of bilingual data: English and the respective language for pretraining. Other works focus on understanding languages LLMs reason in (Wendler et al., 2024), and languages LLMs cannot learn (Borenstein et al., 2024; Kallini et al., 2024). Still, little work has examined how information seen during pretraining in one language can help downstream task performance in another language.

Cross-Lingual Transfer Philippy et al. (2023) present a comprehensive survey on cross-lingual transfer in multilingual language models. The survey explains that cross-lingual transfer is a well studied topic for classic NLP tasks, such as part-of-speech tagging, named entity recognition, dependency parsing, machine translation, etc., although findings are not always consistent. Cross-lingual transfer is less well studied for the language modeling objective, and for modern downstream evaluation tasks, such as ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), SCIQ (Welbl et al., 2017), Wino-

Grande (Sakaguchi et al., 2021), etc. Philippy et al. end with a number of recommendations, one of which is to study cross-lingual transfer in more detail for generative models, given their recent exceptional performance. Our work differs from prior work on cross-lingual transfer in three ways: (i) we do not directly study cross-lingual transfer as we do not first train on the auxiliary language, and then on the target language, (ii) we pretrain 1B and 3B parameters models that are 10-30x the size of typical BERT models, and (iii) instead of finetuning, we focus on how models learn languages during pretraining.

Data Selection Selecting high quality data for pretraining LLMs remains an active area of research. Early research on data selection was based on heuristics. For example, the original GPT-2 model was pretrained on outbound links filtered from Reddit, based on heuristic indicators of whether users found the links interesting, educational, or funny (Radford et al., 2019). Prior approaches also upsampled documents from high quality sources like Wikipedia (Gururangan et al., 2022), or used combinations of heuristics such as absence of stop words, document length, word length, etc. (Rae et al., 2021). Other works select data based on quality filters and time of collection (Longpre et al., 2023), model-based quality filtering (Sachdeva et al., 2024; Li et al., 2024), or textbook quality knowledge (Gunasekar et al., 2023; Li et al., 2023b; Kong et al., 2024). An alternative approach to data selection is re-weighting data samples to select the best data mixtures for training (Fan et al., 2023; Xie et al., 2024), or importance sampling based on a downstream task (Xie et al., 2023; Grangier et al., 2024a,b). Still a majority of these filtering techniques are applied to English-only datasets, and multilingual datasets such as mC4 have limited data filtering (Xue, 2020).

3 Using English Data Selection Pipelines to Complement Limited Target Data

Existing data selection pipelines have been shown to be effective in monolingual (English) pretraining. We investigate whether these pipelines are useful in the bilingual setup with limited target data. In §3.2, we report our findings for experiments with English as the auxiliary language and German as the example target language. In §3.3 we discuss how our findings extend across multiple languages.

3.1 General Implementation Details

Model. We train decoder-only transformer models (Vaswani et al., 2017) with 1.3B parameters. Models use the PolyLM tokenizer (Wei et al., 2023b), with a total vocabulary size of 256K tokens using BPE, covering the unified vocabulary over all languages that we include in our experiments, which allows us to use the same tokenizer across experiments. Models are trained for 100K steps with batch size 1024. Additional hyperparameters and model details are in Appendix A. We report results for the 1.3B model in the main body, and refer to Appendix G for results on a 300M model trained for 30K steps. These model sizes are chosen as they provide reasonable (above random) performance on several benchmark QA tasks, and are commonly used for benchmarking and ablating pretraining of language models (Penedo et al., 2023, 2024).

Data. To meaningfully and systematically study a data constrained scenario, we impose the following trade-off: (i) the number of tokens in the target language needs to be sufficient to start training a language model in the target language (i.e., perplexity needs to go down sufficiently), and (ii) the number of tokens in the target language should be small enough to expect meaningful differences from the auxiliary language. Through empirical exploration, we choose to include 250M tokens in the target language. Not only does this choice satisfy the trade-off posed above, but it also simulates a realistic scenario, as 250M tokens is representative for the amount of data that exists in the tail of mC4 (∼20 languages). For our experiments, we choose languages with more than 250M tokens available in mC4. This allows us to compare with a monolingual setting, to better assess the impact of the auxiliary language. We report additional results for 125M tokens in the target language in Appendix H.

We consider three baselines: (i) *Target (S)*: 250M tokens from the target language, (ii) *Target (L)*: enough target language for 1 training epoch, and (iii) *Auxiliary (L)*: enough auxiliary language for 1 training epoch. For the other experiments, we consider ∼250M tokens from the target language, as stated above, which is repeated to account for 5% of the training steps. The remaining 95% of the training steps consists of the auxiliary data. We refer to the small 250M target language data as "(S)" and to the large (auxiliary) data pool as "(L)".

Evaluation. We consider the average over six general understanding QA tasks: ARC-Easy, ARC-Challenge, Hellaswag, PIQA, SciQ, and Winogrande. These are knowledge-based tasks that small models with limited data still perform well on. Non-English evaluations are conducted via translation of the original dataset for which we use a mix of proprietary large language models. Additional details are provided in §B.2.

3.2 Better English Datasets

Methodology. We first compare the performance of models trained on combinations of German (target) and English (auxiliary) with varying existing English datasets based on the common crawl: mC4 (Xue, 2020), RedPajamav2 (RPJv2) (Computer, 2023), RefinedWeb (RFW) (Penedo et al., 2023), and FineWebEDU (FWE) (Penedo et al., 2024). These datasets have been constructed following a pipeline of different filtering steps outlined in Figure 1a. The resulting datasets are of higher quality and cover different snapshots of the common crawl. Further details are available in Appendix B.

Findings. We report results in Figure 2. For English evaluations, better quality English datasets attain substantially higher performance on the English tasks (up to 9%). For the same benchmarks translated into German, the performance increase is around 2%. Of all compared English datasets, FineWebEDU achieves the best average downstream performance, and within 1% of the DE (L) comparison. There are two primary factors we hypothesize contribute to FineWebEDU achieving better performance on downstream tasks: high-quality data filtering, and relevant information filtering. We investigate this in more detail in §4.

3.3 Experiments Across Multiple Languages

Motivation. We add seven languages, across four language families: French, Italian, Portuguese, Spanish (Indo-European, same as German), Chinese (Sino-Tibetan), Japanese (Japonic), and Korean (Koreanic) (Lewis et al., 2015). We choose these languages because of their variety in language families, the amount of data available in mC4, and the access to translated evaluation data.

Methodology. We train with approximately 250M tokens[1] from the mC4 corpus in the respec-

---

[1] Note that data distributions from the ∼250M data in mC4 looks similar for each language and are not specialized to any particular domain as evidenced by the cluster distribution
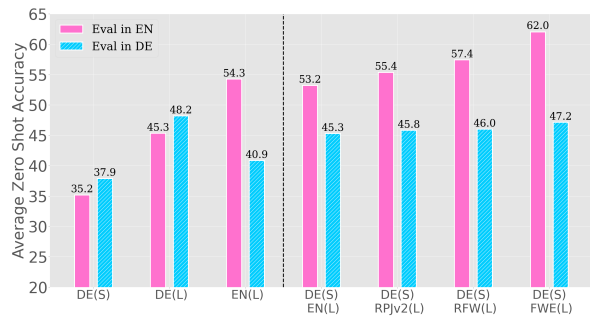

Figure 2: Zero-shot accuracy of models trained with higher quality English auxiliary data. Results are averaged over six eval datasets. We compare training with different auxiliary datasets on English and German evaluations. Better English datasets show large increases in English and smaller increases in German.
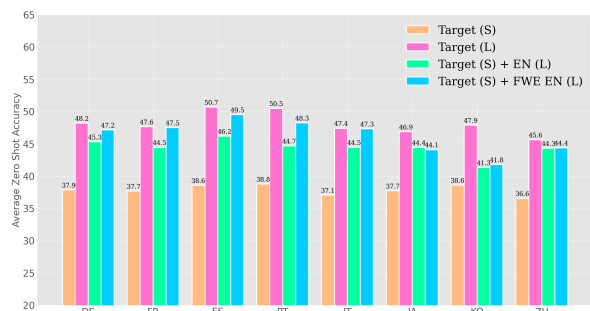

Figure 3: Average zero-shot accuracy in the target language summarized for eight languages. Models trained on 100B tokens. Comparisons between a small and large amount of monolingual data from the target language, a small amount of data from the target language and mC4 English data (same distribution), and a small amount of data from the target language and FineWebEDU.

tive language and denote these models as "target language (S)". The monolingual models are denoted as "target language (L)". The Chinese and Korean mC4 corpora contain fewer than 100B tokens and thus the data is repeated for multiple epochs in the base language experiments. Based on our previous results, we use FineWebEDU as the auxiliary language source. We also conduct additional experiments with Chinese as the auxiliary language and Japanese as the target language, to investigate the effect of linguistic similarity. Although beyond the immediate scope of our investigation, we also investigate a multilingual setup (instead of bilingual) in Appendix I.3.

Findings. We report results for multiple target languages and Chinese auxiliary data in Figure 3-4. We only observe improvements from FineWebEDU for the Indo-European languages in our set, with
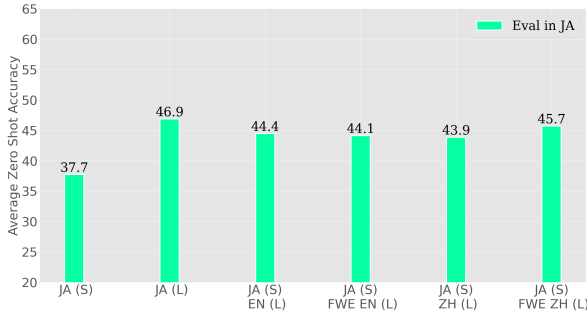
---

visualized in Appendix E.

19099

Figure 4: Average accuracy over zero-shot benchmark tasks in translated Japanese, comparing Chinese and English auxiliary data.
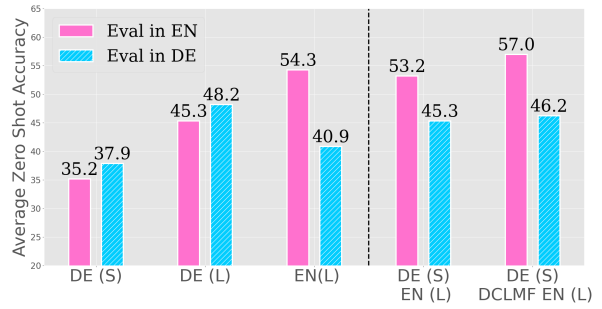

Figure 5: Zero-shot accuracy of models trained with model based filtering of English auxiliary data. Results are averaged over six evaluation datasets. For each setting evaluation is done in English and German.

English as the auxiliary language (Figure 3). Figure 4 shows better performance for Chinese instead of English FineWebEDU as the auxiliary language for Japanese. This is a first indication of the applicability of using different auxiliary languages, based on linguistic similarity.

4 The Effect of Individual Data Transformations

In §3.2 we investigated the effect of using existing data filtering pipelines in the auxiliary language, but this does not isolate which components of the pipeline provide the best downstream improvements. Here, we study individual data transformations that comprise the training pipeline, including model based quality filtering and educational content upsampling, as well as transformations on top of auxiliary data, such as translation. For all experiments, we use the same experimental setup and baselines as in §3, and we focus on German as the target and English as the auxiliary language.

4.1 High Quality Filtering

Motivation.   Prior work shows that model-based filters and higher quality data lead to better performance in English. A key component of FineWeb is custom filters based on text quality (Penedo et al., 2024), and DCLM models are trained on data which has been highly filtered based on heuristic and model-based filters (Li et al., 2024). To isolate the impact high quality auxiliary data can have on target language performance, we apply a recent model-based quality filter on top of mC4 English. We choose this filter as it aims primarily at stylistic quality, and filtering with this model leads to better performance than training with RFW and FWE (Li et al., 2024), two of the best performing datasets in Figure 2. Note that while filtering strategies achieve strong results on English downstream

evaluations, training a filtering model can require more data than available, and high quality datasets may not be available in the target language.

Methodology.   To test the impact of quality filtering, we use the DCLM fast text classifier. This filters data that is aimed at instruction following and high scoring posts in r/ExplainLikeImFive, and was found to be the best quality filter over RFW and other model-based filters (Li et al., 2024). We compare the performance of models trained with English filtering and without, holding the German data the same for evaluation. We refer to this as DCLM Filter (following (Li et al., 2024)) and filter to the top 10% of mC4.

Findings.   The results are in Figure 5. English evaluations improve by 3%. Translated German evaluations are under 1%, within 1 standard error.[2]

4.2 Clustered Dataset Importance Sampling

Motivation.   Prior work shows that LLMs reason in English and that information may be stored in a language agnostic space (Wendler et al., 2024). However, it is unclear whether the relevant information is only seen in English (the predominant language), or also in other languages. Our experiments on FineWebEDU indicate the presence of information sharing, but FineWebEDU filters for both high quality data and for educational quality content (Penedo et al., 2024). In this section we experiment with isolating the information sharing.

Methodology.   To explicitly test whether information is shared between auxiliary and target languages, we upsample topics in English and evaluate on the target language. Given access to some small

---

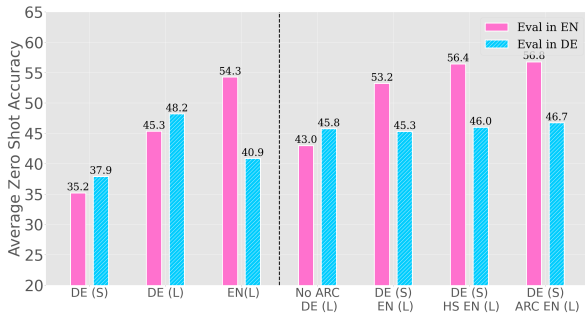[2]Standard errors obtained following (Gao et al., 2024).

Figure 6: Zero-shot accuracy of models trained with upsampling of English auxiliary data. Results are averaged over six evaluation datasets. For each setting evaluation is done in English and German.
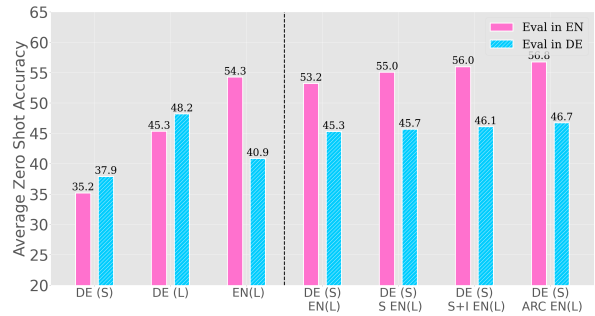

Figure 7: Zero-shot accuracy of models trained with synthetic data upsampling of English auxiliary data. Results are averaged over six evaluation datasets. For each setting evaluation is done in English and German.

target set, importance sampling weights are computed based on the amount of data from a target evaluation set that is assigned to each cluster following (Grangier et al., 2024b).

To train the clustering model, we take a small subset of the training set, produce embeddings from a smaller sentenceBERT model (Reimers and Gurevych, 2020), and cluster the data according to the embeddings. We upweight a subset of roughly 300B tokens from the English dataset. Given a small target set (typically on the order of 1000-10000 samples), we assign each sample to a cluster and upweight the original training set based on the cluster assignment proportions. In practice, we do not optimize for the cluster parameters jointly with the model weights, and instead precompute them based on the pretraining and target task data.

For clustering hyperparameters, we use a lightweight SentenceTransformers multilingual model[3] for extracting features (Reimers and Gurevych, 2019), and a balanced K-means algorithm to cluster the embeddings into 64 clusters.

Upsampling is done with evaluation data only in English, to facilitate having specialized information in the auxiliary language that is unavailable in the target language. We compare two settings: (i) upsampling based on HellaSwag (general knowledge and instruction style), and (ii) upsampling based on ARC Easy (general science knowledge). We also add a comparison where we downsample relevant data to the target task, to compare with having only low quality data, which we call "no ARC DE (L)".

[3]The particular model is called `paraphrase-multilingual-MiniLM-L12-v2` model and is obtained from `https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`.

Findings. Figure 6 shows the results. We see 4% improvement in English evaluations, and 2% improvement in the target language. This highlights that models can take advantage of information in the auxiliary language, and the performance improvements are higher for the information upsampling than for model-based filtering.

In Appendix I.4 we report the results for an additional seven target languages. Generally, our findings indicate that filtering high quality data improves performance in the target language. There is a slight improvement in five of the eight languages for data selection over *relevant topics* but performance overall is similar between filtering strategies and better than no filtering baselines.

4.3 Upsampling with Synthetic Examples

Motivation. While we achieved performance improvements by selecting data based on target downstream evaluations, having such data available at pretraining can be restrictive. For this reason, it can be desirable to be able to *generate* the necessary data for upsampling. While prior work has examined the use of LLM-generated data for pretraining (Maini et al., 2024) and finetuning (Li et al., 2023a; Yuan et al., 2024) language models, to our knowledge there is no prior work that investigates data selection based on synthetic examples.

Methodology We generate a small set of synthetic examples, following the approach in (Maini et al., 2024). The synthetic data is created by prompting an off-the-shelf instruction finetuned language model to generate sets of questions relating to the topic of interest. We use the prompts in Appendix D. Generating synthetic data using an off-the-shelf language model can be both computationally expensive and challenging. However,

19101

for the purpose of computing sampling weights, a small number of questions is sufficient.

For our experiments, we generate data using a frozen Mistral-7B instruction tuned model[4] (Jiang et al., 2023). We create around 2000 science QA pairs while filtering to confirm only a question and answer are specified. We refer to this setting as "S EN (L)", S referring to Science.

We also generate general instruction data, aimed at broad QA style data, and general factual information helpful for downstream tasks. This data is generated in two stages: we use the frozen Mistral-7B model to generate a set of questions, and then the model is prompted to answer the questions.

Using the sets of questions, we identify the cluster sample weights and upsample data accordingly as in §4.2. We refer to this setting as "S+I EN (L)".

Findings We show results in Figure 7. Comparing synthetic data upsampling with upsampling from the downstream tasks (Figure 6) demonstrates that synthetic data can be sufficient for incorporating information into the auxiliary English dataset. Upsampling synthetic data is within 1% for the English downstream tasks, and within 0.5% for the translated German evaluations. Results for transformations in other languages are reported in §I.4.

4.4 Translation Systems

Motivation. Training directly on auxiliary language data can lead to improvements. An alternative strategy is to translate the auxiliary data into the target language, assuming a machine translation system is available. This approach offers the benefit of training the model exclusively in one language, and, if the translation system is of high quality, it allows for training on high-quality data in the target language at the expense of translating the corpus. It is, therefore, important to also assess the level the translation system must possess to effectively translate data for pretraining.

Methodology. For our experiments, we use lightweight translation systems of roughly 100-200M parameters. We consider three models with BLEU scores 16.0, 26.5, and 31.6 on the WMT-17 EN-DE benchmark task. We denote these models as v1, v2, and v3, corresponding to increasing BLEU score. All models are trained on translated versions of the mC4 English corpus. No other English data

_____
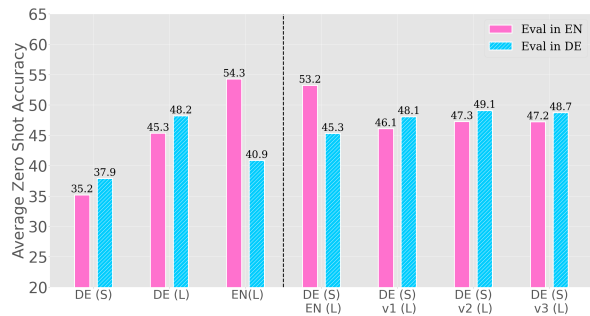[4] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

Figure 8: Zero-shot accuracy of models trained with translations of English auxiliary data. Results are averaged over six evaluation datasets. For each setting evaluation is done in English and German.

is included in the dataset, but we keep the 250M tokens of real German data as in prior experiments.

Findings. Figure 8 shows the results. We find little difference from the quality of the translation. Both stronger translation models achieve similar performance. The v1 model performs comparably to training on real German data, and the data translated by the v2 and v3 models obtain around 1% improvement in comparison to real German data.

We hypothesize a few reasons this may be possible, but leave investigation of each of these to future work: (1) English CC data is of higher quality than German CC data, as there may be more data from more diverse sources. (2) Translated German data has a different distribution from real German data, and this better matches the translated test evaluations. (3) Translated data from a small translation system might simplify language, which makes it easier for models to learn, following (Eldan and Li, 2023). (4) Portions of the dataset could not be translated by the systems and are removed. These portions might be noisy, and some unintended filtering may lead to slightly higher performance.

5 Data and Model Scaling

To study the impact of data and model scaling, we investigate two scaling questions: (1) How does the required quantity of additional auxiliary data relate to the amount of data required if one had enough target language data to train on? (§5.1) (2) How do findings extend to a larger bilingual model? (§5.2) Following our setup in the previous section, we target our investigation to German (target) and English (auxiliary).
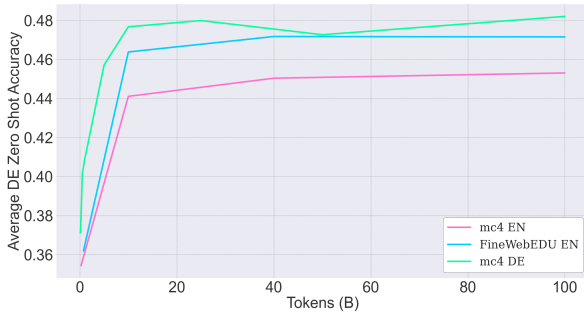
Figure 9: Average accuracy over zero-shot benchmark tasks in translated German, with increasing number of tokens in both target and auxiliary languages. Models are trained for 100B tokens.
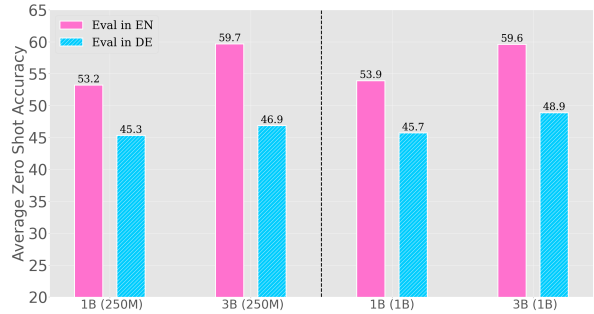

Figure 10: Zero-shot accuracy for 1B and 3B models trained with 250M or 1B tokens in the target language denoted in brackets. Models are trained on the target language data for different ratios of the training steps . See §F for details on data ratios.

5.1 Auxiliary Data Scaling

Motivation. We measure the amount of data needed in the *target language* to match training on high quality auxiliary data. Our goal is to quantify the advantage of training models with additional target data, beyond the 250M tokens used before.

Methodology. We investigate data scaling over mC4 and FineWebEDU. We train models at different dataset sizes, from 0.1B to 100B tokens.

Findings. Our results are summarized in Figure 9. Models trained on the English mC4 data achieve similar performance to models trained with only 5B tokens of German mC4 data. In contrast, training on a high quality data such as FineWebEDU increases the amount of German data needed to around 10B tokens, 2x the amount of German data. Note that the curves all plateau quickly at around 10B tokens, corresponding to around 10 repetitions of the data, matching (Muennighoff et al., 2024).

There are two important data scaling considerations from Figure 9 that justify training with auxiliary language data. First, for languages that are data constrained, it may be infeasible to collect twice as much data. Second, models trained on FineWebEDU attain similar performance at a rate of 5x the number of tokens (roughly 50B tokens of FineWebEDU matches the performance of 10B tokens of mC4 German data). An important avenue is to investigate data scaling at larger quantities of tokens. In particular, the FineWebEDU corpus totals 5.4T tokens and would require access to 1T tokens of German data, which is 3x the amount of data in mC4. As a result, while the data scaling shows large improvements from little German data, the large amount of readily available English data can make training on auxiliary data practical.

5.2 Model Size Scaling

Next, we investigate to what extent results scale for training larger language models when data remains constrained in the target language. Given a fixed amount of data from the target language (250M tokens) training a larger model may be impractical, as, according to Chinchilla scaling laws (Kaplan et al., 2020), increasing model size necessitates increasing the number of tokens seen during training. This is challenging for data constrained languages, as the number of repetitions increases with increased amount of training. This can lead to overfitting and saturated model performance.

We train a ∼2.7B parameter model for 150K steps, matching the same scaling ratio as for the 1.3B model. We train models following the DE (S) + EN (L) setup. Our main results after experimenting with multiple data ratios and sizes are in Figure 10.

To demonstrate that training larger models for data constrained models may be infeasible, and illustrate the performance differences at different scales in Figure 10, we first show the perplexity of training and validation data in Figure 11 for the 1.3B (left) and 2.7B models (right). For 1.3B models, we see little to no overfitting at 5% ratio of training steps, and the model achieves slightly lower perplexity than a model trained with a lower ratio of 1.5% target language data. In contrast, for the 2.7B model, there is clear overfitting from as early as 25% of the training steps with 5% ratio of training steps for the data constrained language. This indicates that the number of repetitions is too high for the 2.7B model and performance may degrade.

To show this, we evaluate both 1.3B and 2.7B models with 5% data ratios with 250M tokens in
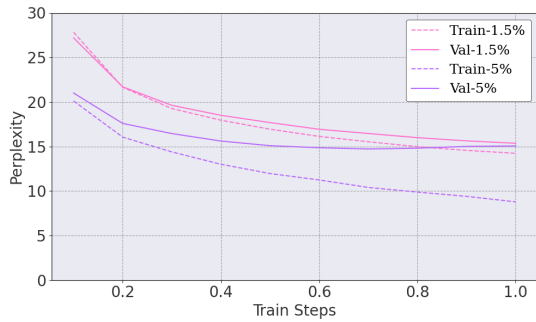
(a) 1B Model Perplexity
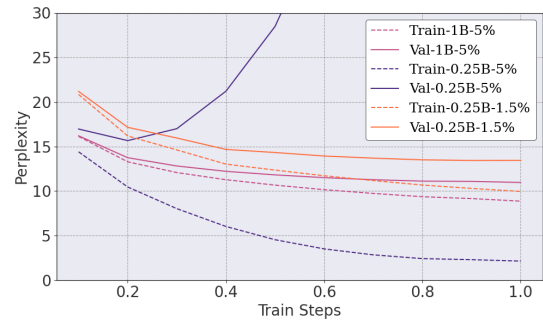
(b) 3B Model Perplexity

Figure 11: Train and validation perplexity for 1.3B and 2.7B parameters with varying amount of data.

(a) 250M Tokens

(b) 1B Tokens

Figure 12: Average accuracy over zero-shot benchmark tasks in English and translated German comparing 1.3B and 2.7B models trained with 5% of the training steps using (a) 250M tokens, and (b) 1B tokens in German.
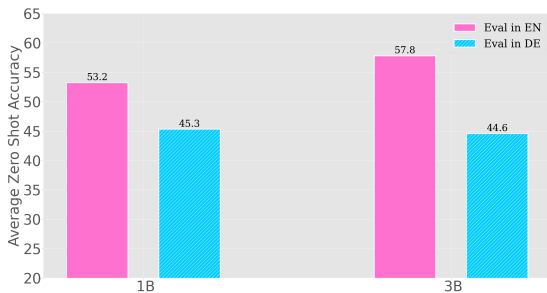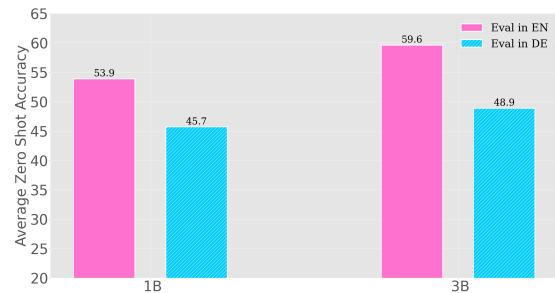
the target language. Results in Figure 12 show that the 2.7B model performs slightly worse than the 1.3B model with the same 250M amount of target language data matching the similar perplexity values between the two models. However, increasing to 1B tokens results in larger improvements when increasing model size from 1.3B to 2.7B, as shown in Figure 12b.

We further conduct a comparison with different data ratios. We experiment with different data ratios beyond 5% used during training. For this experiment, we study the 1.3B parameter model and 2.7B parameter model. Zero-shot accuracy is provided in Figure 17 for all data ratios in the appendix. We see that performance is around the same for 1.3B models regardless of ratio, but 2.7B models perform better with lower data ratios. Our final results training 2.7B models with reduced proportion of steps are in Figure 10.

In summary, there is a limit to the size of models that can be trained with limited target data. Model performance for 250M tokens (more than available in mC4 for ~20 languages) shows limited improvement when scaling up from a 1.3B to a 2.7B parameter model. In contrast, increasing to 1B tokens doubles improvement when increasing model size

to a 2.7B parameter model. Training a 2.7B parameter model necessitates training on at least 1B tokens of data to see larger gains in performance (more than available in mC4 for ~40 languages).

6 Conclusion

This work studies how an auxiliary language for which an abundance of training data is available can boost pretraining for a target language for which only limited data is available. We find that adding auxiliary high quality data obtained by data filtering can improve performance in a target language. However, we find that results are inconsistent across target languages. We hypothesize that for languages further from English, better English datasets are not as helpful as information is not shared between them. Finally, we find limitations to scaling models for languages that are data constrained. This work takes a step towards pretraining language models in languages with limited data, and can inspire more research into bilingual learning under data constraints.

7 Limitations

In this section we list some limitations of our work.

Evaluation data. When evaluating language models for languages other than English, one often needs to rely on translated evaluation sets. Not only may this introduce translation mistakes, the resulting evaluation set might also contain cultural biases. As a result, certain aspects of the evaluation may lead to improved performance when using English auxiliary or translated data. Additionally, translated data often exhibits a distribution different from that of real data in the target languages. Therefore, an important direction for future work is the development of evaluation datasets that are not based on translation, which is essential for more accurate evaluation of bi- and multilingual language models.

Languages included. The focus of this work is on training language models for data constrained languages. We note that there are many languages within mC4 (and more broadly) which can benefit from having auxiliary English data for pretraining. Due to the aforementioned limited evaluation benchmarks and availability of target language data for comparison, we leave investigation for these languages to future work.

Acknowledgements

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

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

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

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

Nadav Borenstein, Anej Svete, Robin Chan, Josef Valvoda, Franz Nowak, Isabelle Augenstein, Eleanor Chodroff, and Ryan Cotterell. 2024. What languages are easy to language-model? a perspective from learning probabilistic regular languages. *arXiv preprint arXiv:2406.04289*.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

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

Tyler A Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K Bergen. 2023. When is multilinguality a curse? language modeling for 250 high-and low-resource languages. *arXiv preprint arXiv:2311.09205*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Together Computer. 2023. Redpajama: an open dataset for training large language models.

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

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from pos tagging with over 100 languages. In *The 60th Annual Meeting of the Association for*

Computational Linguistics, pages 7676–7685. Association for Computational Linguistics (ACL).

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Fahim Faisal and Antonios Anastasopoulos. 2024. An efficient approach for studying cross-lingual transfer in multilingual language models. *arXiv preprint arXiv:2403.20088*.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. 2023. Doge: Domain reweighting with generalization estimation. *arXiv preprint arXiv:2310.15393*.

Manuel Faysse, Patrick Fernandes, Nuno Guerreiro, António Loison, Duarte Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro Martins, et al. 2024. Croissantllm: A truly bilingual french-english language model. *arXiv preprint arXiv:2402.00786*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

David Grangier, Simin Fan, Skyler Seto, and Pierre Ablin. 2024a. Task-adaptive pretrained language models via clustered-importance sampling. *arXiv preprint arXiv:2410.03735*.

David Grangier, Angelos Katharopoulos, Pierre Ablin, and Awni Hannun. 2024b. Specialized language models with cheap inference from limited domain data. *arXiv preprint arXiv:2402.01093*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Suchin Gururangan, Dallas Card, Sarah K Dreier, Emily K Gade, Leroy Z Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A Smith. 2022. Whose language counts as high quality? measuring language ideologies in text data selection. *arXiv preprint arXiv:2201.10474*.

Yotam Intrator, Matan Halfon, Roman Goldenberg, Reut Tsarfaty, Matan Eyal, Ehud Rivlin, Yossi Matias, and Natalia Aizenberg. 2024. Breaking the language barrier: Can direct inference outperform pre-translation in multilingual LLM applications? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 829–844, Mexico City, Mexico. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. *arXiv preprint arXiv:2401.06416*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Xiang Kong, Tom Gunter, and Ruoming Pang. 2024. Large language model-guided document selection. *arXiv preprint arXiv:2406.04638*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2019. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2015. Ethnologue: Languages of the world, eighteenth edition. SIL International, Dallas, Texas.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. 2024. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. 2023a. Self-alignment with instruction back-translation. *arXiv preprint arXiv:2308.06259*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023b. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

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

Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. *arXiv preprint arXiv:2401.16380*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.

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

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

T Pires. 2019. How multilingual is multilingual bert. *arXiv preprint arXiv:1906.01502*.

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

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

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. 2024. How to train data-efficient llms. *arXiv preprint arXiv:2402.09668*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Raphael Scheible, Fabian Thomczyk, Patric Tippmann, Victor Jaravine, and Martin Boeker. 2020. Gottbert: a pure german language model. *arXiv preprint arXiv:2012.02110*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, et al. 2023a. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023b. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.

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. 2024. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. 2023. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36:34201–34227.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A survey on multilingual large language models: Corpora, alignment, and bias. *arXiv preprint arXiv:2404.00929*.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

A Hyperparameters and Training Details

The medium-scale (300M non-embedding parameter) model consists of 24 layers, 16 attention heads, and a hidden dimension size of 1024. The XL-scale (1.3B non-embedding parameter) model consists of 24 layers, 16 attention heads, and a hidden dimension size of 2048. Both models have a maximum sequence length of 1024.

The baseline models are trained using NVIDIA's Megatron-LM[5] repository for pretraining language models. The medium size models are trained for a total of 30K steps, and 100K steps for the XL models at a batch size of 1024. All models are trained using a maximum learning rate of 0.0003 for the medium model and 0.0002 for the XL model, and a minimum learning rate of 0.00001 with a cosine learning rate scheduler and warmup for 1% of the total steps. For regularization, we use a weight decay of 0.01, along with a gradient clipping norm of 1.0. Models are trained with the Adam optimizer using $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

The 2.7B parameter model consists of 32 layers with 2560 hidden dimension and 32 attention heads. 2.7B parameter models are trained for a total of 150K steps with a batch size of 1024 and context length of 2048. All models are trained using a learning rate of 0.00016 with the same optimization recipe as other models.

The total training time for 2.7B parameter models for 100K steps is roughly 3000 GPUh on Nvidia H100 GPUs. XL models trained on roughly 100K steps are around 1000 GPUh. For medium size models, the total training time is around 200 hours for roughly 30K steps.

B Dataset Details

B.1 Train Sets

- **mC4**: The primary pretraining corpus in our experiments is multilingual Colossal Clean Crawled Corpus (mC4), a curated text dataset comprising over 6.3T tokens. This corpus is derived from CommonCrawl and used for pretraining numerous language models (Brown et al., 2020; Raffel et al., 2020; Touvron et al., 2023). The dataset is chosen as all languages have similar data extraction pipelines including line length filter, cld3 language detection, and deduplication (Xue, 2020). The English portion contains 2.7T tokens, German contains 350B tokens,

French contains 320B tokens, Spanish contains 430B tokens, Portuguese contains 146B tokens, Italian contains 160B tokens, Korean contains 26B tokens, Japanese contains 160B tokens, and Chinese contains 40B tokens. For both Chinese and Korean monolingual models: Target (L), we repeat data up to 100K steps of training (roughly 2-3 repetitions).

- **RedPajamav2**: A pretraining corpus with light filtering (primarily only deduplication) comprising 30T tokens and 20T tokens of English text. We focus on the English portion of the dataset only and train using a random shuffled subset of both the head and middle portions (Computer, 2023).

- **RefinedWeb**: The dataset is also derived from the CommonCrawl, however has a more stringent filtering process including trafilatura text extraction, document and line level rules, and fuzzy duplication over the original C4 processing (Penedo et al., 2023).

- **FineWeb**: This dataset is derived from the CommonCrawl with the aim of replicating RefinedWeb at larger scales. The dataset has some additional filtering including Gopher filtering (Rae et al., 2021), additional C4 filters, and custom filters for text quality (Penedo et al., 2024).

- **FineWebEDU**: A subset of the FineWeb dataset which is filtered according to a classifier trained on annotations for educational quality from Llama-3 70B model (Penedo et al., 2024).

- **ChineseFineWeb-EDU**[6]: An educational corpus in Chinese consisting of roughly 400B tokens of data. Although it shares a similar name, the ChineseFineWeb-EDU does not share data from FineWebEDU and is collected from different sources. Models trained on this data are trained with only 1 repetition.

Note that while the original datasets aim to be entirely monolingual, we do observe incidental data in other languages. While we expect the amount of incidental data to be small, it is difficult to quantify and demonstrate its effect directly. We note that prior work by Blevins and Zettlemoyer (2022) shows that around 0.05% of the T5 dataset contains non-English data, amounting to millions of tokens

---

[5] `https://github.com/NVIDIA/Megatron-LM`

[6] `https://huggingface.co/datasets/opencsg/chinese-fineweb-edu-v2`

in other languages. This is due to errors in LangID methods, which are shown to not be robust to scale of web text data (Caswell et al., 2020). In mC4, the LangID filtering that is used in many state of the art filtering techniques, and although we cannot be fully certain, it seems likely to assume that each setting that we compare is somewhat equally affected by incidental target data.

B.2 Zero Shot Evaluations

- **SciQ**: A dataset of science exam questions, specifically designed to evaluate the ability of NLP models in understanding and reasoning within the scientific domain (Welbl et al., 2017).

- **ARC Challenge (ARC-C)**: This dataset is part of the AI2 Reasoning Challenge (ARC) (Clark et al., 2018), containing science exam questions from grades 3 to 9. The ARC Challenge set includes more difficult questions that necessitate higher-order reasoning.

- **ARC Easy (ARC-E)**: The Easy set of the AI2 Reasoning Challenge (Clark et al., 2018) features questions from the same source as ARC-C but are considered less challenging and do not require as advanced reasoning skills.

- **Winogrande (Wino.)**: This dataset challenges models on common sense reasoning in a language context, focusing on pronoun disambiguation tasks (Sakaguchi et al., 2021).

- **PIQA**: Physical Interaction Question Answering tests the understanding of everyday physical processes, an aspect of practical common sense (Bisk et al., 2020).

- **HellaSwag**: This dataset evaluates a model's ability to complete scenarios in a contextually and logically coherent manner, requiring both language understanding and common sense reasoning (Zellers et al., 2019).

For each of the eval datasets, we include the number of samples for each translated evaluation in Table 1. For our evaluations, we use the lm-eval-harness repository[7] for zero-shot accuracy on QA tasks.

We translate primarily using our own systems to (i) ensure that translation artifacts will be consistent across evaluation tasks, (ii) have translations for all

---

[7] https://github.com/EleutherAI/lm-evaluation-harness

languages we evaluate. We note further that evaluation frameworks such as Okapi (Dac Lai et al., 2023) do not release evaluation for all languages we study, and only support HellaSwag as they do not translate the full ARC evaluation set. We did not evaluate on other multilingual evaluation tasks such as MT and MGSM as these are not traditionally knowledge tasks that require information sharing between the two languages.

B.3 Number of Data Files for Filtering Experiments

The mC4 English portion of the dataset is split into roughly 11,264 files totaling 2.7T tokens of data (Xue, 2020). For each of our experiments, data is filtered differently, and as such varying numbers of files are needed for training. At a baseline, we consider the first 1500 files totaling roughly 350B tokens of data. This number was selected to match the total amount of German data which is recorded as 347B tokens using the mT5 tokenizer (Xue, 2020). For the OH classifier, we use the first 10,000 files and filter down to 10% of the dataset. For German, Japanese, Spanish, Portuguese, Italian, and French models, we use the first two files of data totaling roughly 250-300M tokens of data. For Chinese and Korean models, we use the first 7 files totaling roughly 250M tokens.

B.4 License and Attribution

All datasets used in this paper are supported by public licenses including ODC and Apache. The pre-trained models including Mistral and OH FastText classifiers are also supported by public licenses including Apache and MIT licenses. We use the Megatron codebase under the Nvidia license for pre-training and the lm-eval-harness (MIT) for evaluations. All models and datasets are collected from Huggingface via the datasets library where possible. We use a proprietary translation system for fast translation at scale and are thus unable to provide details of the license at this time.

C Evaluation Metrics

The metric utilized for evaluation is the *macro token level perplexity*. Given a batch of encoded texts, the perplexity at the token level was computed as follows:

Given the accumulated loss over the entire dataset, denoted as L, and the total number of tokens, represented by T, the macro token-level perplexity, denoted as $\mathcal{P}$, is calculated as:

Dataset	EN	DE	FR	ZH	JA	PT	ES	IT	KO
ARC-C	1,172	1,137	1,147	1,146	1,147	1,117	1,117	1,117	1,147
ARC-E	2,376	2,260	2,271	2,271	2,271	2,271	2,271	2,271	2,271
HS	10,042	9,368	9,338	9,266	10,033	9,229	9,374	9,193	10,025
PIQA	1,838	1,838	1,838	1,838	1,838	1,838	1,838	1,838	1,838
SCIQ	1,000	950	953	1,000	926	953	951	952	945
WG	1,267	1,184	1,215	1,059	1,096	1,232	1,239	1,233	1,177

Table 1: Evaluation set sizes for each language.

$$\mathcal{P} = \exp\left(\min\left(20, \frac{L}{T}\right)\right) \quad (1)$$

Where:

- $\exp$ is the exponential function.
- L is the cumulative loss over all shifted logits and labels in the dataset.
- T is the total number of tokens in the dataset.

The value of 20 acts as an upper limit to stabilize the metric in cases of high loss values.

For zero-shot MCQ accuracy evaluations, we compute the perplexity of each sentence completion, and choose the lowest perplexity choice. We use the lm-evaluation-harness and where possible evaluate with the length-normalized accuracy. Unless otherwise stated, all evaluations are zero-shot.

D Synthetic Prompts and Examples

For building the synthetic corpus used in our data selection experiments, we consider three prompts for generating science questions (similar to many downstream tasks), fact-based QA data, and instruction-based writing (such as emails, books, lists, etc.). For generating science questions, we generate both the question and answer. For the fact and instruction data, we first generate the questions using the prompt, and subsequently generate the answer without any additional prompting.

Science Question Prompt

```
Give me a set of ten question and
answer pairs on topics relating to
Physics, Chemistry and Biology that
a high school student would be able
to answer. The response should be
in the form Question: <question>
\n Answer: <answer> \n \n with an
answer that is less than ten words.
The response should not contain any
other details or explanations about
the question or answer.
```

Facts Question Prompt

```
People from different social
and educational backgrounds,
beliefs, ethnicity and gender
are asking an AI assistant for
information. They are looking
for detailed explanations about
encyclopedic facts on Wikipedia
and in textbooks, about philosophy,
nature, science, entertainment,
literature, geography, sociology,
law, history, etc. Write an
interesting and difficult question
that would be sent to the AI
assistant:
```

Instruction Writing Prompt

```
People from different social and
educational backgrounds, beliefs,
ethnicity and gender are asking an
AI assistant to help them write a
piece of text that they need for
their work or their personal life.
They can ask the AI Assistant to
write a document (email, letter,
official document...). Each request
comes with a long, precise and
detailed description of what needs
to be in the text, and why they
need this document.  The request
may also include information about
the writing style, the tone, the
target audience or the layout of
the text.  The description of the
task is formal, detailed and clear.
Each request is composed of a few
paragraphs written in English, and
starts with the tag <request>. Here
is some of the most interesting and
original requests sent to the AI
assistant:
```

E Data Distributions

We present results comparing the data distributions for each dataset within the mC4 English clusters. The purpose of these plots is to illustrate that the data from other languages follows a similar distribution to that of mC4 English. Figure 13 shows the distribution for French and German languages, Spanish and Portuguese in Figure 14, Italian and Korean in Figure 15, and for Japanese and Chinese languages in Figure 16.

F Model Size Data Scaling

In this section, we include additional findings accompanying the results presented in §5.2. Figure 17 show performance of 1.3B and 2.7B models at different ratios of data for the target language.

G Results for 300M Models

We present results comparing approaches for 300M parameter models with English as the auxiliary language and German as the target language.

Figure 18a shows performance with different datasets, Figure 18b shows performance with the DCLM filter, Figure 18c shows performance with upsampling real data, Figure 18d shows performance upsampling synthetic data, and Figure 18e shows performance with translation. All results show similar trends as in Section 3-4, but the performance improvements are smaller.

H Results for Smaller Target Language Data Size

We conduct an experiment at higher number of repetitions (34) with only 125M tokens in German, and using English as the auxiliary language. 125M tokens corresponds to around one file of mC4 German data, and is similar to the lowest data sizes in mC4; there are only six out of 107 languages that have fewer than 100M tokens. The results are shown in Table 2. We refer to the 125M German corpus as DE (XS) and to the 250M German tokens as DE (S). Results for DE(XS) + EN(L) are 5.5% higher than 125M tokens of German data only (DE XS), which is similar to the increase of DE (S) + EN (L) over 250M tokens of German data (DE (S)) (which is 7%). This is a first indication that our method also works for settings with fewer tokens available in the target language.

I Results for Multiple Languages

I.1 Perplexity Evaluations for Translated Training Data

In §3.3 we found that performance trends were not the same across languages. In particular, French, German, Portuguese, and Spanish (belonging to the same language family) have similar patterns, but Chinese, Japanese, and Korean exhibit different patterns. To further test whether the models retain knowledge from one language in another, we translate a small portion of the training set from FineWebEDU and mC4 English, totaling 10,000 documents. We then translate the data using the v3 translation system from §4.4. We measure both the macro perplexity of all documents as well as the fraction of times where the translated and original data from FineWebEDU (training set) have lower loss than the average loss of documents from mC4 English (not part of the training set but from a similar distribution). We refer to this quantity as translated and original exceedance. Having lower loss means the data is more familiar to the model, and having an equal exceedance across the original and translated data means the model can reason equally in either language. Our results are summarized
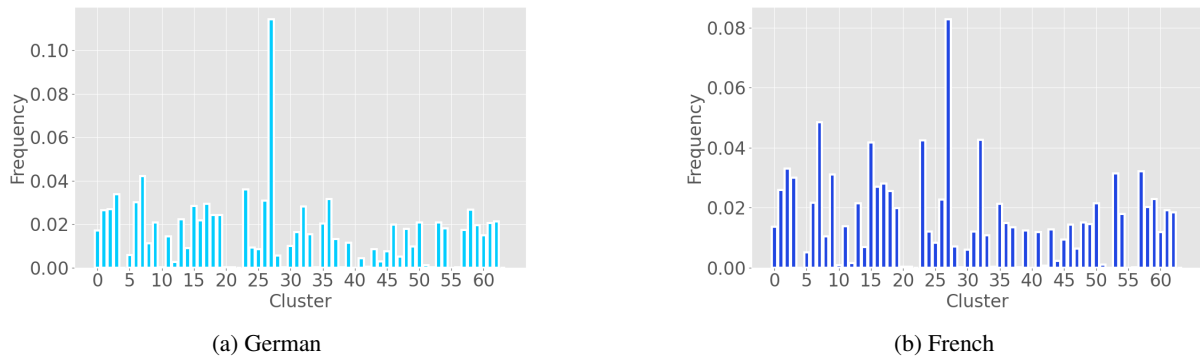
(a) German

(b) French

Figure 13: Data distribution within mC4 English clusters for German and French mC4 data.
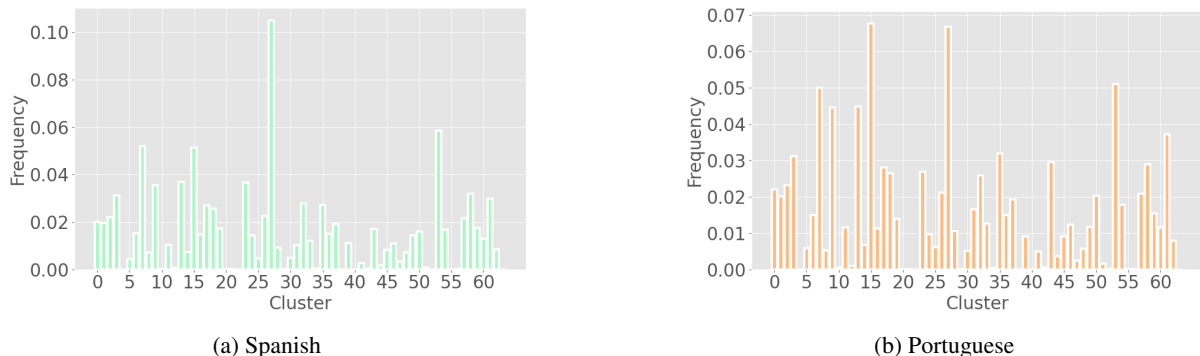

(a) Spanish

(b) Portuguese

Figure 14: Data distribution within mC4 English clusters for Spanish and Portuguese mC4 data.

in Table 3. We find that perplexity is nearly identical for original data, but much higher for translated data in all languages. For exceedance, in English, we see that the scores are all around 80%. However, we see that for Japanese and Chinese these values are much lower, indicating that seeing the data in English for these languages does not lower the perplexity in the target language and that the model is not making use of information in the other language. For Chinese evaluations, we note that the perplexity is much higher than for other languages, indicating that the translation system potentially causes higher perplexity and lower exceedance. However, we still note that for Japanese the exceedance is lower, and we expect that with better translation quality, the Chinese evaluations will be similar to Japanese.

I.2 Other Auxiliary Languages

In this section we give additional details for our experiments with other auxiliary languages than English, as presented in §3.3.

Methodology. We conduct experiments comparing using English and Chinese as an auxiliary language with Japanese as the target language. For these experiments, we train a 1.3B model

with the auxiliary dataset being either mC4 in English or Chinese, and English or Chinese versions of FineWebEDU[8]. Our evaluations are on the Japanese translations.

Findings. We summarize the results in Figure 19. Our findings indicate that using FineWebEDU, a high quality Chinese dataset, as auxiliary data improves Japanese performance over mC4 in Chinese, and the performance increases over using English either from mC4 or FineWebEDU. Note that the performance increases are consistent with improvements better English data for Indo-European languages as seen in Figure 3.

I.3 Multilingual Experiments

Motivation. Our main findings investigate how auxiliary data benefits evaluations in a target language. However, many models trained on other languages (beyond English) are trained with many languages simultaneously. We investigate whether better auxiliary language datasets also improve multilingual model training.

---

[8] https://huggingface.co/datasets/opencsg/chinese-fineweb-edu-v2

19113
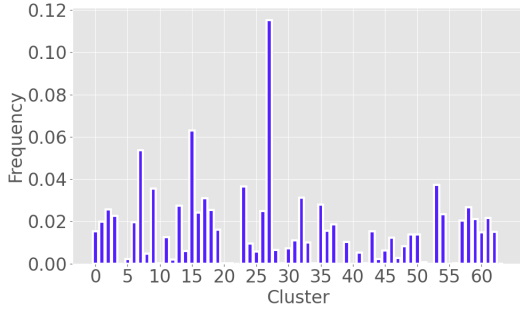
(a) Italian

(b) Korean

Figure 15: Data distribution within mC4 English clusters for Italian and Korean mC4 data.

(a) Japanese
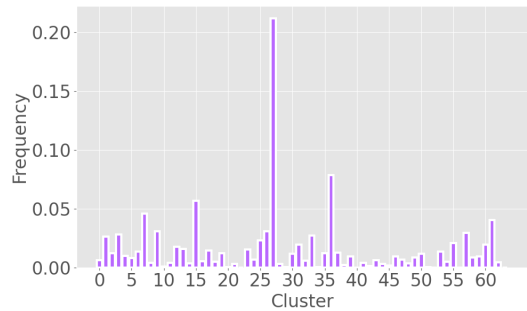
(b) Chinese

Figure 16: Data distribution within mC4 English clusters for Japanese and Chinese mC4 data.
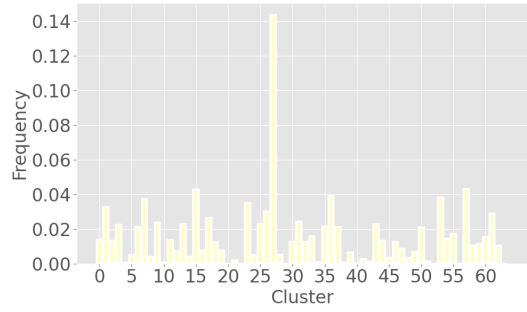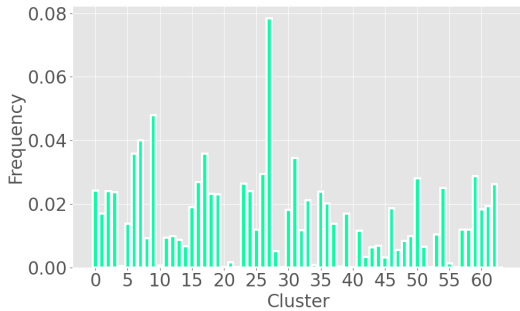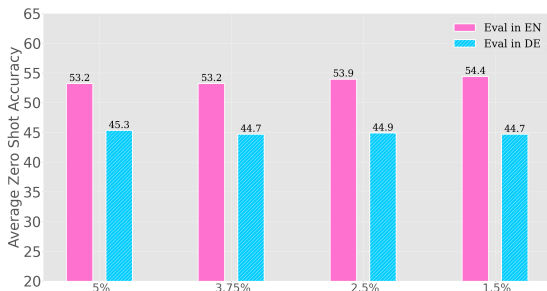

(a) 1.3B Model
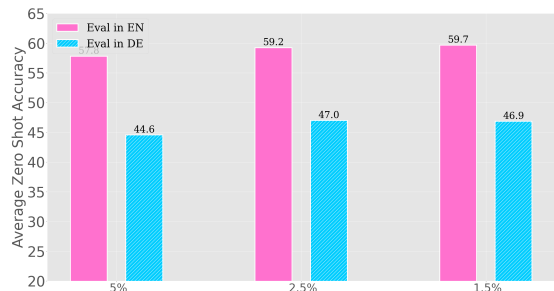
(b) 2.7B Model

Figure 17: Zero-shot QA performance for 1B and 3B models at varying data ratios during training. For all models the total amount of available data in the target language is 250M tokens.

Methodology. We conduct experiments combining the German, French, Chinese, and Japanese language data towards multilingual training. For these experiments, we train a 1.3B model with the auxiliary dataset being FineWebEDU and a mix of data from the four languages, totaling 5% or 20% of the training. We use this subset of four languages, as we keep the data ratios the same per language, and did not want to increase the amount of data in target languages beyond the typical multilingual ratios in large open-source models (Xu et al., 2024).

Findings. We summarize the results in Table 4. Our findings indicate that training with 20% of the

data being a combination of target languages yields similar performance to training with 5%, resulting in a 1% reduction in performance on average when training with 20% multilingual data. When compared with training a bilingual model, we observe performance decreases for German and French, and increases for Chinese and Japanese.

I.4 Average Zero Shot Accuracy Plots

We present experimental results comparing the best performing approaches for French and German languages in Figure 20, Spanish and Portuguese in Figure 21, Italian and Korean in Figure 22, and for Japanese and Chinese languages in Figure 23. For
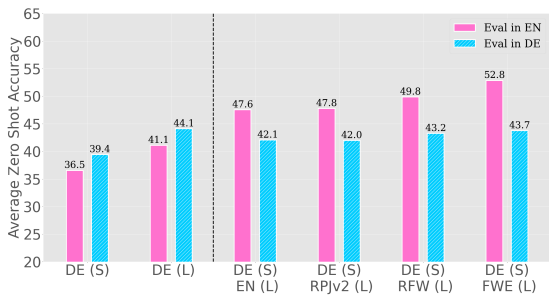
(a) Higher quality English auxiliary data.
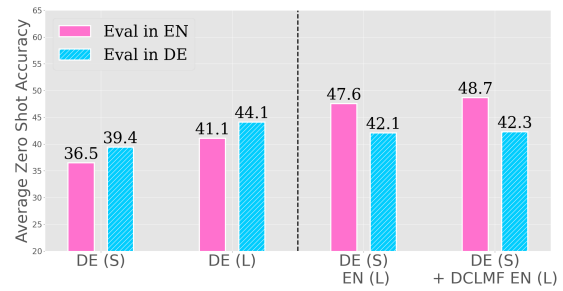
(b) Model based filtering of English auxiliary data.
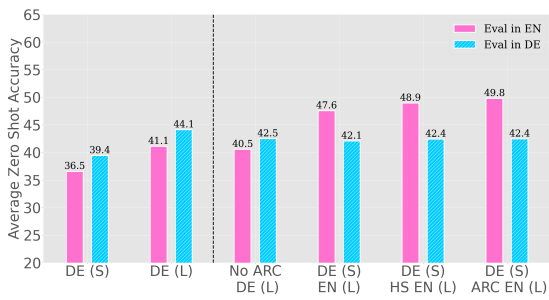
(c) Upsampling of English auxiliary data.
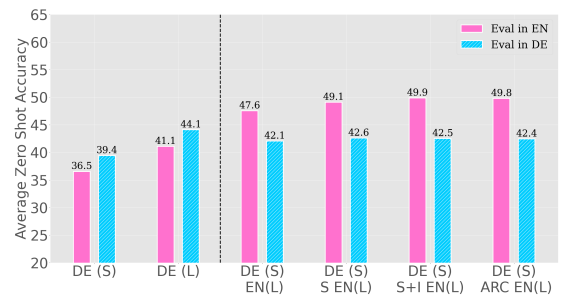
(d) Synthetic data upsampling of English auxiliary data.
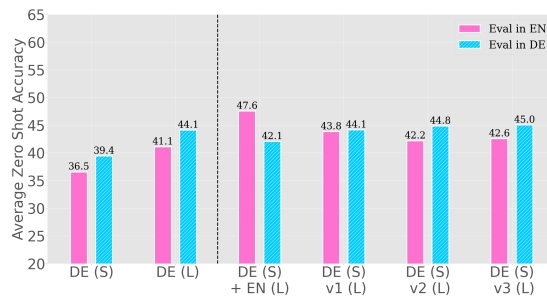
(e) Translations of English auxiliary data.

Figure 18: Zero-shot accuracy of 300M models trained with different types of English auxiliary data. Results are averaged over six evaluation datasets. For each setting evaluation is done in English (pink, left bars) and German (blue, right bars).

Model Name	ARC-C-DE	ARC-E-DE	HS-DE	PIQA-DE	SCIQ-DE	WGrande-DE	AVG-DE
DE (XS)	23.66 ± 1.26	30.13 ± 0.97	27.19 ± 0.46	54.13 ± 1.16	37.68 ± 1.57	49.92 ± 1.45	37.12
DE(XS) + EN(L)	24.27 ± 1.27	33.63 ± 0.99	32.96 ± 0.49	56.80 ± 1.16	55.05 ± 1.61	53.21 ± 1.45	42.65
DE (S)	22.69 ± 1.24	30.31 ± 0.97	28.99 ± 0.47	54.84 ± 1.16	40.00 ± 1.59	50.42 ± 1.45	37.88
DE (S) + EN (L)	25.07 ± 1.29	37.70 ± 1.02	36.18 ± 0.50	59.36 ± 1.15	63.05 ± 1.57	50.51 ± 1.45	45.31

Table 2: Performance comparison for 125M tokens of target data vs. 250M tokens of target data.

Language	mC4-Train	mc4-Val	mC4-EN	mC4-EN Translated	FWE	FWE Translated	Original EX	Translated EX
German	8.41	16.41	14.25	25.58	10.61	21.31	78.70	75.00
French	6.24	12.75	14.37	20.31	10.66	14.37	79.24	88.64
Japanese	6.52	11.21	14.56	25.56	10.64	23.89	80.41	56.97
Chinese	4.35	21.38	15.37	165.90	10.69	210.43	84.07	27.71

Table 3: Perplexity evaluations for mC4 English and FineWeb-EDU comparing original data and translated versions for 1B models trained with 250M tokens from the target language and FineWeb-EDU as the auxiliary dataset.
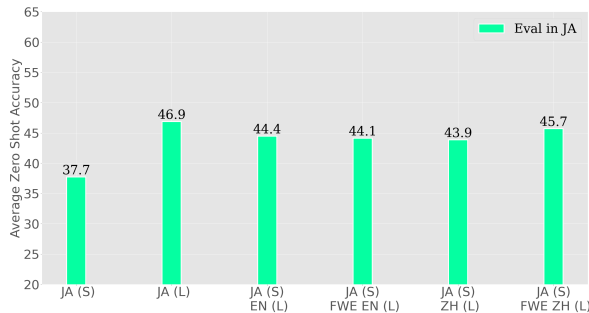
Figure 19: Average accuracy over zero-shot benchmark tasks in translated Japanese comparing Chinese and English auxiliary data. Models are 1.3B size and trained for 100B tokens.

	EN	DE	FR	JA	ZH
5%	61.07	46.02	46.64	**44.00**	**45.92**
20%	59.25	46.44	46.61	43.23	44.36
Bi		**47.16**	**47.52**	42.73	44.38

Table 4: Evaluation of XL models in multilingual setting on "General Understanding Tasks" focusing on general reasoning, language understanding, and science knowledge in translated languages. Rows are the average accuracy for the respective language, with 5% or 20% of the training coming from a mix of the four languages. 'Bi' refers to the bilingual models trained with target (S) + FWE EN (L).

each language, we take the best performing dataset and model found in German from §3.2-3.3.

I.5 Individual Eval Dataset Results

Results for individual evaluation datasets are shown for all languages in Tables 5-12.
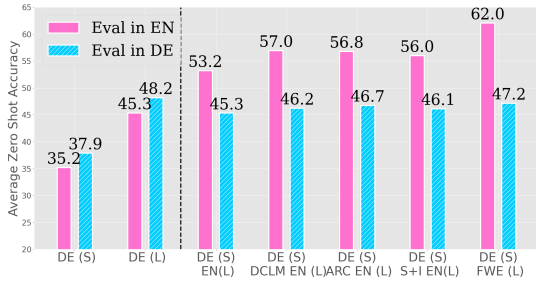
(a) German
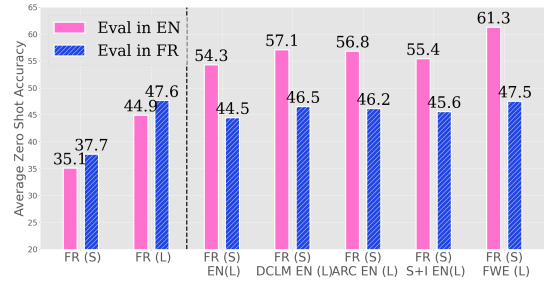
(b) French

Figure 20: Zero-shot accuracy of XL models trained with various English auxiliary data for German and French. Results are averaged over six eval datasets.
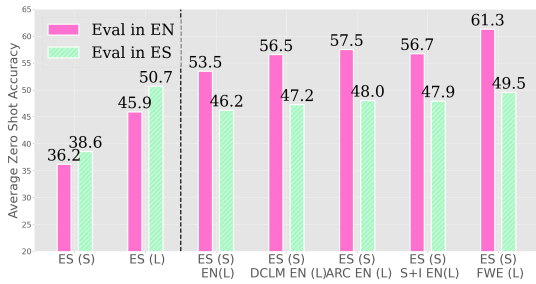
(a) Spanish
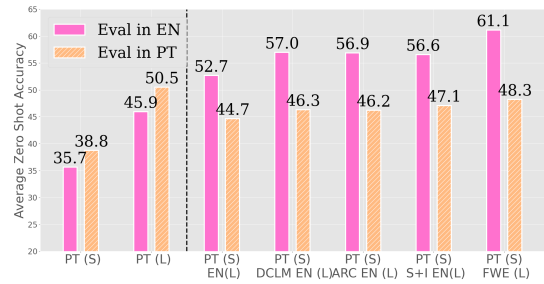
(b) Portuguese

Figure 21: Zero-shot accuracy of XL models trained with various English auxiliary data for Spanish and Portuguese. Results are averaged over six eval datasets.
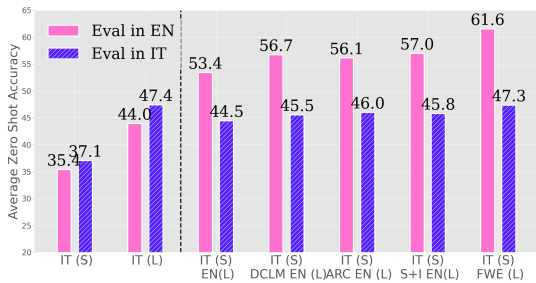
(a) Italian
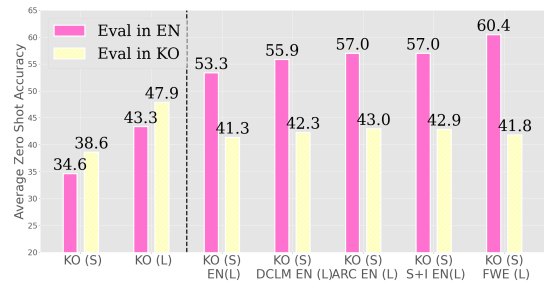
(b) Korean

Figure 22: Zero-shot accuracy of XL models trained with various English auxiliary data for Italian and Korean. Results are averaged over six eval datasets.
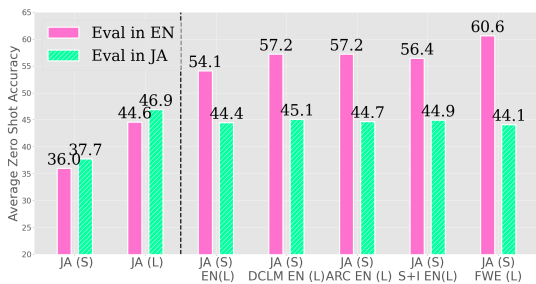
(a) Japanese
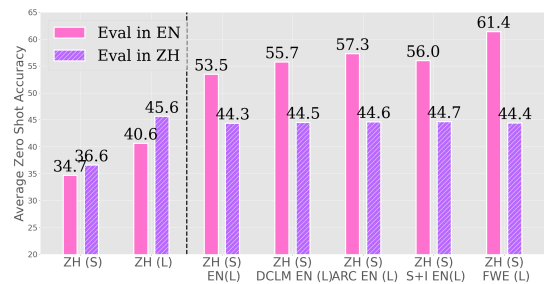
(b) Chinese

Figure 23: Zero-shot accuracy of XL models trained with various English auxiliary data for Japanese and Chinese. Results are averaged over six eval datasets.

J Results for Individual datasets for German Target Language Models

Results for the 300M models on individual eval datasets are also provides in Tables 13-14. Results for 1B models on English evaluation tasks are shown in Table 15 and for translated German benchmarks in Table 16.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WGrande	AVG
DE (S)	20.73 ± 1.18	27.53 ± 0.92	25.94 ± 0.44	52.88 ± 1.16	36.80 ± 1.53	47.12 ± 1.40	35.17
DE (L)	22.78 ± 1.23	37.92 ± 1.00	33.06 ± 0.47	62.35 ± 1.13	64.70 ± 1.51	51.14 ± 1.40	45.33
DE (S) EN(L)	25.94 ± 1.28	48.15 ± 1.03	48.50 ± 0.50	71.55 ± 1.05	73.10 ± 1.40	52.01 ± 1.40	53.21
DE (S) DCLM EN (L)	31.06 ± 1.35	57.24 ± 1.02	49.26 ± 0.50	71.71 ± 1.05	78.40 ± 1.30	54.06 ± 1.40	56.96
DE (S) ARC EN (L)	31.83 ± 1.36	57.24 ± 1.02	48.80 ± 0.50	73.07 ± 1.04	77.00 ± 1.33	52.72 ± 1.40	56.78
DE (S) S+I EN(L)	29.10 ± 1.33	55.22 ± 1.02	48.34 ± 0.50	71.71 ± 1.05	78.50 ± 1.30	53.04 ± 1.40	55.98
DE (S) FWE (L)	38.14 ± 1.42	66.37 ± 0.97	54.88 ± 0.50	72.25 ± 1.04	84.60 ± 1.14	56.04 ± 1.39	62.05
DE (S)	22.69 ± 1.24	30.31 ± 0.97	28.99 ± 0.47	54.84 ± 1.16	40.00 ± 1.59	50.42 ± 1.45	37.88
DE (L)	27.44 ± 1.32	38.81 ± 1.03	39.53 ± 0.51	63.17 ± 1.13	67.68 ± 1.52	52.62 ± 1.45	48.21
DE (S) EN(L)	25.07 ± 1.29	37.70 ± 1.02	36.18 ± 0.50	59.36 ± 1.15	63.05 ± 1.57	50.51 ± 1.45	45.31
DE (S) DCLM EN (L)	26.21 ± 1.30	40.09 ± 1.03	35.86 ± 0.50	59.30 ± 1.15	64.00 ± 1.56	51.94 ± 1.45	46.23
DE (S) ARC EN (L)	26.91 ± 1.32	39.34 ± 1.03	35.77 ± 0.50	58.65 ± 1.15	67.58 ± 1.52	52.11 ± 1.45	46.73
DE (S) S+I EN(L)	25.24 ± 1.29	39.38 ± 1.03	35.34 ± 0.49	59.74 ± 1.14	64.21 ± 1.56	52.70 ± 1.45	46.10
DE (S) FWE (L)	26.91 ± 1.32	42.39 ± 1.04	37.40 ± 0.50	60.45 ± 1.14	65.37 ± 1.54	50.42 ± 1.45	47.16

Table 5: Evaluation of 1B parameter XL model on "General Understanding Tasks" focusing on general reasoning, language understanding, and science knowledge in English followed by translated German. Results show the length normalized accuracy for individual datasets and the average over all datasets for all datasets.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WGrande	AVG
FR (S)	19.88 ± 1.17	26.60 ± 0.91	26.61 ± 0.44	52.56 ± 1.17	33.70 ± 1.50	51.14 ± 1.40	35.08
FR (L)	24.66 ± 1.26	36.78 ± 0.99	32.89 ± 0.47	60.66 ± 1.14	63.90 ± 1.52	50.51 ± 1.41	44.90
FR (S) EN(L)	26.28 ± 1.29	49.92 ± 1.03	48.62 ± 0.50	71.98 ± 1.05	75.20 ± 1.37	53.91 ± 1.40	54.32
FR (S) DCLM EN (L)	30.12 ± 1.34	56.52 ± 1.02	49.22 ± 0.50	71.71 ± 1.05	80.30 ± 1.26	54.78 ± 1.40	57.11
FR (S) ARC EN (L)	29.69 ± 1.34	57.28 ± 1.02	48.96 ± 0.50	72.91 ± 1.04	78.90 ± 1.29	53.12 ± 1.40	56.81
FR (S) S+I EN(L)	28.41 ± 1.32	53.16 ± 1.02	47.44 ± 0.50	70.51 ± 1.06	78.90 ± 1.29	54.14 ± 1.40	55.43
FR (S) FWE (L)	36.69 ± 1.41	65.40 ± 0.98	54.81 ± 0.50	72.74 ± 1.04	82.50 ± 1.20	55.41 ± 1.40	61.26
FR (S)	22.06 ± 1.22	28.53 ± 0.95	29.25 ± 0.47	55.93 ± 1.16	40.29 ± 1.59	50.04 ± 1.44	37.68
FR (L)	25.98 ± 1.30	38.53 ± 1.02	41.71 ± 0.51	64.25 ± 1.12	62.12 ± 1.57	53.25 ± 1.43	47.64
FR (S) EN(L)	23.54 ± 1.25	37.03 ± 1.01	37.66 ± 0.50	58.98 ± 1.15	60.44 ± 1.58	49.05 ± 1.43	44.45
FR (S) DCLM EN (L)	24.93 ± 1.28	38.62 ± 1.02	38.49 ± 0.50	60.45 ± 1.14	64.53 ± 1.55	52.02 ± 1.43	46.51
FR (S) ARC EN (L)	25.54 ± 1.29	39.32 ± 1.03	38.17 ± 0.50	59.41 ± 1.15	64.22 ± 1.55	50.29 ± 1.44	46.16
FR (S) S+I EN(L)	25.72 ± 1.29	38.71 ± 1.02	37.61 ± 0.50	59.03 ± 1.15	63.06 ± 1.56	49.55 ± 1.43	45.61
FR (S) FWE (L)	27.38 ± 1.32	40.51 ± 1.03	40.67 ± 0.51	60.94 ± 1.14	64.85 ± 1.55	50.78 ± 1.43	47.52

Table 6: Evaluation of 1B parameter XL model on "General Understanding Tasks" focusing on general reasoning, language understanding, and science knowledge in English followed by French. Results show the length normalized accuracy for individual datasets and the average over all datasets for all datasets.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WGrande	AVG
ES (S)	21.16 ± 1.19	28.45 ± 0.93	26.33 ± 0.44	52.61 ± 1.16	38.50 ± 1.54	50.04 ± 1.41	36.18
ES (L)	23.29 ± 1.24	38.55 ± 1.00	33.86 ± 0.47	60.83 ± 1.14	67.90 ± 1.48	51.07 ± 1.40	45.92
ES (S) EN(L)	26.19 ± 1.28	48.95 ± 1.03	48.16 ± 0.50	71.11 ± 1.06	74.30 ± 1.38	52.01 ± 1.40	53.45
ES (S) DCLM EN (L)	30.46 ± 1.34	56.36 ± 1.02	49.48 ± 0.50	71.38 ± 1.05	78.40 ± 1.30	53.20 ± 1.40	56.55
ES (S) ARC EN (L)	30.89 ± 1.35	58.71 ± 1.01	49.41 ± 0.50	73.45 ± 1.03	78.50 ± 1.30	54.06 ± 1.40	57.50
ES (S) S+I EN(L)	29.44 ± 1.33	57.49 ± 1.01	48.89 ± 0.50	70.57 ± 1.06	79.70 ± 1.27	54.22 ± 1.40	56.72
ES (S) FWE (L)	36.60 ± 1.41	64.86 ± 0.98	54.95 ± 0.50	71.87 ± 1.05	81.50 ± 1.23	57.85 ± 1.39	61.27
ES (S)	21.80 ± 1.22	29.90 ± 0.96	29.84 ± 0.47	57.45 ± 1.15	41.85 ± 1.60	50.61 ± 1.42	38.57
ES (L)	27.99 ± 1.33	44.12 ± 1.04	43.64 ± 0.51	65.94 ± 1.11	69.72 ± 1.49	52.70 ± 1.42	50.69
ES (S) EN(L)	25.28 ± 1.28	38.71 ± 1.02	39.32 ± 0.50	60.94 ± 1.14	65.83 ± 1.54	47.13 ± 1.42	46.20
ES (S) DCLM EN (L)	26.94 ± 1.31	40.82 ± 1.03	40.15 ± 0.51	60.39 ± 1.14	65.83 ± 1.54	49.31 ± 1.42	47.24
ES (S) ARC EN (L)	25.89 ± 1.29	42.01 ± 1.04	39.49 ± 0.50	62.02 ± 1.13	66.25 ± 1.53	52.54 ± 1.42	48.03
ES (S) S+I EN(L)	25.89 ± 1.29	41.08 ± 1.03	39.35 ± 0.50	61.37 ± 1.14	67.93 ± 1.51	51.82 ± 1.42	47.91
ES (S) FWE (L)	27.99 ± 1.33	43.24 ± 1.04	42.05 ± 0.51	62.19 ± 1.13	70.45 ± 1.48	51.17 ± 1.42	49.51

Table 7: Evaluation of 1B parameter XL model on "General Understanding Tasks" focusing on general reasoning, language understanding, and science knowledge in English followed by Spanish. Results show the length normalized accuracy for individual datasets and the average over all datasets for all datasets.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WGrande	AVG
PT (S)	23.38 ± 1.24	26.77 ± 0.91	26.01 ± 0.44	53.48 ± 1.16	33.50 ± 1.49	50.91 ± 1.41	35.67
PT (L)	23.98 ± 1.25	39.14 ± 1.00	33.26 ± 0.47	60.99 ± 1.14	66.30 ± 1.50	52.01 ± 1.40	45.95
PT (S) EN(L)	25.34 ± 1.27	48.02 ± 1.03	47.19 ± 0.50	71.60 ± 1.05	74.00 ± 1.39	50.04 ± 1.41	52.70
PT (S) DCLM EN (L)	28.92 ± 1.33	57.70 ± 1.01	48.78 ± 0.50	71.87 ± 1.05	80.50 ± 1.25	54.14 ± 1.40	56.99
PT (S) ARC EN (L)	30.72 ± 1.35	57.83 ± 1.01	48.79 ± 0.50	73.01 ± 1.04	78.30 ± 1.30	52.64 ± 1.40	56.88
PT (S) S+I EN(L)	29.35 ± 1.33	57.45 ± 1.01	48.26 ± 0.50	71.55 ± 1.05	78.80 ± 1.29	53.99 ± 1.40	56.56
PT (S) FWE (L)	35.07 ± 1.39	64.39 ± 0.98	54.81 ± 0.50	72.69 ± 1.04	82.10 ± 1.21	57.62 ± 1.39	61.11
PT (S)	25.37 ± 1.29	30.74 ± 0.97	29.18 ± 0.47	55.39 ± 1.16	41.97 ± 1.60	50.00 ± 1.43	38.77
PT (L)	30.25 ± 1.36	43.55 ± 1.04	42.36 ± 0.51	64.80 ± 1.11	69.88 ± 1.49	52.11 ± 1.42	50.49
PT (S) EN(L)	24.06 ± 1.26	38.04 ± 1.02	36.04 ± 0.50	59.68 ± 1.14	59.18 ± 1.59	51.06 ± 1.42	44.68
PT (S) DCLM EN (L)	27.29 ± 1.32	39.85 ± 1.03	37.63 ± 0.50	59.58 ± 1.14	63.27 ± 1.56	50.41 ± 1.43	46.34
PT (S) ARC EN (L)	27.90 ± 1.32	38.79 ± 1.02	36.87 ± 0.50	60.07 ± 1.14	64.32 ± 1.55	49.35 ± 1.42	46.22
PT (S) S+I EN(L)	27.03 ± 1.31	40.38 ± 1.03	37.72 ± 0.50	60.72 ± 1.14	66.00 ± 1.54	50.73 ± 1.42	47.10
PT (S) FWE (L)	28.86 ± 1.34	43.11 ± 1.04	39.81 ± 0.51	60.34 ± 1.14	66.84 ± 1.53	50.57 ± 1.42	48.25

Table 8: Evaluation of 1B parameter XL model on "General Understanding Tasks" focusing on general reasoning, language understanding, and science knowledge in English followed by Portuguese. Results show the length normalized accuracy for individual datasets and the average over all datasets for all datasets.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WGrande	AVG
IT (S)	23.46 ± 1.24	27.23 ± 0.91	26.30 ± 0.44	53.21 ± 1.16	32.40 ± 1.48	49.88 ± 1.41	35.41
IT (L)	21.59 ± 1.20	35.77 ± 0.98	32.54 ± 0.47	59.09 ± 1.15	63.70 ± 1.52	51.07 ± 1.40	43.96
IT (S) EN(L)	26.11 ± 1.28	48.53 ± 1.03	47.16 ± 0.50	71.71 ± 1.05	73.50 ± 1.40	53.35 ± 1.40	53.39
IT (S) DCLM EN (L)	29.35 ± 1.33	56.90 ± 1.02	48.99 ± 0.50	71.98 ± 1.05	79.20 ± 1.28	53.83 ± 1.40	56.71
IT (S) ARC EN (L)	30.46 ± 1.34	56.06 ± 1.02	48.78 ± 0.50	73.01 ± 1.04	75.90 ± 1.35	52.57 ± 1.40	56.13
IT (S) S+I EN(L)	30.38 ± 1.34	56.78 ± 1.02	48.49 ± 0.50	71.27 ± 1.06	80.20 ± 1.26	54.85 ± 1.40	56.99
IT (S) FWE (L)	37.03 ± 1.41	65.61 ± 0.97	54.91 ± 0.50	72.74 ± 1.04	84.20 ± 1.15	54.85 ± 1.40	61.56
IT (S)	23.63 ± 1.25	29.19 ± 0.95	28.49 ± 0.47	56.31 ± 1.16	36.76 ± 1.56	48.01 ± 1.42	37.07
IT (L)	26.33 ± 1.30	40.38 ± 1.03	39.89 ± 0.51	64.74 ± 1.11	61.76 ± 1.58	51.42 ± 1.42	47.42
IT (S) EN(L)	25.81 ± 1.29	36.02 ± 1.01	35.29 ± 0.50	58.65 ± 1.15	58.82 ± 1.60	52.23 ± 1.42	44.47
IT (S) DCLM EN (L)	25.81 ± 1.29	39.06 ± 1.02	36.65 ± 0.50	60.66 ± 1.14	62.29 ± 1.57	48.82 ± 1.42	45.55
IT (S) ARC EN (L)	25.72 ± 1.29	39.15 ± 1.02	36.18 ± 0.50	60.55 ± 1.14	63.24 ± 1.56	51.34 ± 1.42	46.03
IT (S) S+I EN(L)	26.07 ± 1.30	40.03 ± 1.03	36.67 ± 0.50	58.76 ± 1.15	62.29 ± 1.57	50.93 ± 1.42	45.79
IT (S) FWE (L)	29.90 ± 1.35	41.04 ± 1.03	38.24 ± 0.51	62.68 ± 1.13	61.03 ± 1.58	51.18 ± 1.42	47.34

Table 9: Evaluation of 1B parameter XL model on "General Understanding Tasks" focusing on general reasoning, language understanding, and science knowledge in English followed by Italian. Results show the length normalized accuracy for individual datasets and the average over all datasets for all datasets.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WGrande	AVG
KO (S)	23.04 ± 1.23	26.14 ± 0.90	25.41 ± 0.43	50.98 ± 1.17	29.50 ± 1.44	52.57 ± 1.40	34.61
KO (L)	22.10 ± 1.21	37.46 ± 0.99	28.99 ± 0.45	59.19 ± 1.15	60.70 ± 1.55	51.62 ± 1.40	43.34
KO (S) EN(L)	26.02 ± 1.28	47.47 ± 1.02	47.71 ± 0.50	71.11 ± 1.06	75.00 ± 1.37	52.72 ± 1.40	53.34
KO (S) DCLM EN (L)	29.69 ± 1.34	55.39 ± 1.02	49.53 ± 0.50	71.65 ± 1.05	75.80 ± 1.36	53.12 ± 1.40	55.86
KO (S) ARC EN (L)	30.03 ± 1.34	57.37 ± 1.01	49.41 ± 0.50	73.78 ± 1.03	78.20 ± 1.31	53.12 ± 1.40	56.98
KO (S) S+I EN(L)	30.97 ± 1.35	58.54 ± 1.01	48.16 ± 0.50	71.49 ± 1.05	80.30 ± 1.26	52.33 ± 1.40	56.97
KO (S) FWE (L)	36.01 ± 1.40	63.59 ± 0.99	54.41 ± 0.50	72.96 ± 1.04	80.10 ± 1.26	55.56 ± 1.40	60.44
KO (S)	24.41 ± 1.27	28.84 ± 0.95	27.63 ± 0.45	52.39 ± 1.17	47.41 ± 1.63	50.89 ± 1.46	38.60
KO (L)	28.07 ± 1.33	42.18 ± 1.04	35.48 ± 0.48	60.66 ± 1.14	71.64 ± 1.47	49.28 ± 1.46	47.89
KO (S) EN(L)	22.76 ± 1.24	33.82 ± 0.99	28.88 ± 0.45	55.22 ± 1.16	56.08 ± 1.62	51.23 ± 1.46	41.33
KO (S) DCLM EN (L)	24.59 ± 1.27	32.54 ± 0.98	29.69 ± 0.46	53.16 ± 1.16	63.17 ± 1.57	50.64 ± 1.46	42.30
KO (S) ARC EN (L)	25.28 ± 1.28	33.73 ± 0.99	29.33 ± 0.45	53.86 ± 1.16	64.44 ± 1.56	51.15 ± 1.46	42.97
KO (S) S+I EN(L)	23.19 ± 1.25	34.26 ± 1.00	30.01 ± 0.46	55.60 ± 1.16	64.02 ± 1.56	50.04 ± 1.46	42.86
KO (S) FWE (L)	22.93 ± 1.24	32.32 ± 0.98	29.09 ± 0.45	53.26 ± 1.16	64.76 ± 1.55	48.43 ± 1.46	41.80

Table 10: Evaluation of 1B parameter XL model on "General Understanding Tasks" focusing on general reasoning, language understanding, and science knowledge in English followed by Korean. Results show the length normalized accuracy for individual datasets and the average over all datasets for all datasets.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WGrande	AVG
JA (S)	22.61 ± 1.22	28.49 ± 0.93	26.29 ± 0.44	53.21 ± 1.16	33.80 ± 1.50	51.46 ± 1.40	35.98
JA (L)	23.46 ± 1.24	37.42 ± 0.99	29.07 ± 0.45	59.14 ± 1.15	68.10 ± 1.47	50.28 ± 1.41	44.58
JA (S) EN(L)	26.19 ± 1.28	49.24 ± 1.03	48.69 ± 0.50	72.63 ± 1.04	74.40 ± 1.38	53.28 ± 1.40	54.07
JA (S) DCLM EN (L)	29.86 ± 1.34	58.29 ± 1.01	48.93 ± 0.50	72.20 ± 1.05	80.20 ± 1.26	53.67 ± 1.40	57.19
JA (S) ARC EN (L)	29.35 ± 1.33	56.94 ± 1.02	49.05 ± 0.50	73.78 ± 1.03	79.80 ± 1.27	54.22 ± 1.40	57.19
JA (S) S+I EN(L)	30.38 ± 1.34	56.14 ± 1.02	48.32 ± 0.50	72.42 ± 1.04	78.60 ± 1.30	52.64 ± 1.40	56.42
JA (S) FWE EN (L)	36.09 ± 1.40	63.93 ± 0.99	54.74 ± 0.50	73.18 ± 1.03	80.00 ± 1.27	55.80 ± 1.40	60.62
JA (S)	22.58 ± 1.24	30.07 ± 0.96	29.32 ± 0.45	53.81 ± 1.16	41.79 ± 1.62	48.81 ± 1.51	37.73
JA (L)	25.28 ± 1.28	40.69 ± 1.03	35.11 ± 0.48	58.98 ± 1.15	69.98 ± 1.51	51.19 ± 1.51	46.87
JA (S) EN(L)	25.28 ± 1.28	36.15 ± 1.01	31.71 ± 0.46	56.26 ± 1.16	66.41 ± 1.55	50.82 ± 1.51	44.44
JA (S) DCLM EN (L)	27.46 ± 1.32	37.78 ± 1.02	31.71 ± 0.46	57.07 ± 1.15	66.63 ± 1.55	49.82 ± 1.51	45.08
JA (S) ARC EN (L)	27.55 ± 1.32	37.43 ± 1.02	31.56 ± 0.46	57.29 ± 1.15	65.55 ± 1.56	48.72 ± 1.51	44.68
JA (S) S+I EN(L)	27.38 ± 1.32	36.42 ± 1.01	31.44 ± 0.46	56.58 ± 1.16	67.28 ± 1.54	50.36 ± 1.51	44.91
JA (S) FWE EN (L)	25.89 ± 1.29	35.71 ± 1.01	31.06 ± 0.46	56.20 ± 1.16	66.09 ± 1.56	49.64 ± 1.51	44.10

Table 11: Evaluation of 1B parameter XL model on "General Understanding Tasks" focusing on general reasoning, language understanding, and science knowledge in English followed by translated Japanese. Results show the length normalized accuracy for individual datasets and the average over all datasets for all datasets.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WGrande	AVG
ZH (S)	22.10 ± 1.21	26.14 ± 0.90	25.74 ± 0.44	52.50 ± 1.17	31.10 ± 1.46	50.43 ± 1.41	34.67
ZH (L)	21.16 ± 1.19	33.16 ± 0.97	27.63 ± 0.45	55.98 ± 1.16	56.60 ± 1.57	49.09 ± 1.41	40.61
ZH (S) EN(L)	25.85 ± 1.28	47.90 ± 1.03	48.54 ± 0.50	71.93 ± 1.05	73.90 ± 1.39	52.64 ± 1.40	53.46
ZH (S) DCLM EN (L)	29.35 ± 1.33	57.24 ± 1.02	48.28 ± 0.50	70.95 ± 1.06	77.20 ± 1.33	51.38 ± 1.40	55.73
ZH (S) ARC EN (L)	30.72 ± 1.35	57.49 ± 1.01	48.38 ± 0.50	73.29 ± 1.03	80.10 ± 1.26	53.75 ± 1.40	57.29
ZH (S) S+I EN(L)	30.38 ± 1.34	56.31 ± 1.02	47.66 ± 0.50	71.16 ± 1.06	78.10 ± 1.31	52.49 ± 1.40	56.02
ZH (S) FWE (L)	36.18 ± 1.40	67.13 ± 0.96	54.07 ± 0.50	73.99 ± 1.02	80.90 ± 1.24	55.96 ± 1.40	61.37
ZH (S)	22.77 ± 1.24	28.58 ± 0.95	28.46 ± 0.47	50.71 ± 1.17	40.80 ± 1.55	48.06 ± 1.54	36.56
ZH (L)	25.65 ± 1.29	38.88 ± 1.02	33.07 ± 0.49	56.37 ± 1.16	69.90 ± 1.45	49.86 ± 1.54	45.62
ZH (S) EN(L)	25.22 ± 1.28	36.77 ± 1.01	31.62 ± 0.48	56.09 ± 1.16	68.00 ± 1.48	48.35 ± 1.54	44.34
ZH (S) DCLM EN (L)	23.30 ± 1.25	36.72 ± 1.01	32.58 ± 0.49	53.92 ± 1.16	69.70 ± 1.45	50.80 ± 1.54	44.50
ZH (S) ARC EN (L)	23.21 ± 1.25	37.52 ± 1.02	32.06 ± 0.48	55.44 ± 1.16	70.90 ± 1.44	48.54 ± 1.54	44.61
ZH (S) S+I EN(L)	22.16 ± 1.23	38.00 ± 1.02	32.24 ± 0.49	54.03 ± 1.16	68.60 ± 1.47	53.07 ± 1.53	44.68
ZH (S) FWE (L)	25.04 ± 1.28	36.50 ± 1.01	31.89 ± 0.48	54.62 ± 1.16	66.30 ± 1.50	51.94 ± 1.54	44.38

Table 12: Evaluation of 1B parameter XL model on "General Understanding Tasks" focusing on general reasoning, language understanding, and science knowledge in English followed by translated Chinese. Results show the length normalized accuracy for individual datasets and the average over all datasets for all datasets.

Small DE	21.33 ± 1.20	27.90 ± 0.92	26.20 ± 0.44	51.20 ± 1.17	42.70 ± 1.56	49.72 ± 1.41	36.51
No ARC Large DE	21.08 ± 1.19	31.52 ± 0.95	28.09 ± 0.45	56.47 ± 1.16	54.80 ± 1.57	51.30 ± 1.40	40.54
Large DE	20.73 ± 1.18	33.33 ± 0.97	28.15 ± 0.45	57.40 ± 1.15	56.30 ± 1.57	50.59 ± 1.41	41.09
Small DE + Large EN	23.81 ± 1.24	41.92 ± 1.01	35.73 ± 0.48	67.03 ± 1.10	66.10 ± 1.50	50.75 ± 1.41	47.56
Small DE + DCLM Filter EN (L)	24.57 ± 1.26	45.20 ± 1.02	35.27 ± 0.48	65.94 ± 1.11	69.00 ± 1.46	52.01 ± 1.40	48.67
Small DE + Large HS EN	23.72 ± 1.24	42.42 ± 1.01	39.64 ± 0.49	70.13 ± 1.07	67.20 ± 1.49	50.43 ± 1.41	48.93
Small DE +Large ARC EN	26.28 ± 1.29	46.80 ± 1.02	36.44 ± 0.48	67.46 ± 1.09	70.30 ± 1.45	51.38 ± 1.40	49.78
+ Large S EN	26.62 ± 1.29	49.33 ± 1.03	31.94 ± 0.47	63.44 ± 1.12	72.40 ± 1.41	50.83 ± 1.41	49.09
Small DE + Large S+I EN	26.19 ± 1.28	46.93 ± 1.02	36.02 ± 0.48	66.21 ± 1.10	73.10 ± 1.40	50.67 ± 1.41	49.85
Small DE + Large v1	21.16 ± 1.19	36.83 ± 0.99	29.19 ± 0.45	60.07 ± 1.14	63.50 ± 1.52	52.33 ± 1.40	43.84
Small DE + Large v2	21.16 ± 1.19	34.89 ± 0.98	29.60 ± 0.46	57.45 ± 1.15	59.50 ± 1.55	50.36 ± 1.41	42.16
Small DE + Large v3	19.54 ± 1.16	35.02 ± 0.98	29.46 ± 0.45	59.47 ± 1.15	61.40 ± 1.54	50.51 ± 1.41	42.57
RPJv2	25.09 ± 1.27	43.27 ± 1.02	37.23 ± 0.48	65.02 ± 1.11	66.30 ± 1.50	49.80 ± 1.41	47.78
RefinedWeb	24.40 ± 1.26	43.98 ± 1.02	39.75 ± 0.49	68.66 ± 1.08	69.80 ± 1.45	52.49 ± 1.40	49.85
FineWeb	25.00 ± 1.27	44.23 ± 1.02	40.89 ± 0.49	69.53 ± 1.07	68.40 ± 1.47	51.78 ± 1.40	49.97
FineWebEDU	28.67 ± 1.32	56.06 ± 1.02	40.85 ± 0.49	66.65 ± 1.10	72.60 ± 1.41	52.09 ± 1.40	52.82

Table 13: Evaluation of 300M parameter medium model on "General Understanding Tasks" focusing on general reasoning, language understanding, and science knowledge in English. Results show the length normalized accuracy for individual datasets and the average over all datasets for all datasets.

Model Name	ARC-C-DE	ARC-E-DE	HS-DE	PIQA-DE	SCIQ-DE	WGrande-DE	AVG-DE
Small DE	23.83 ± 1.26	30.71 ± 0.97	28.33 ± 0.47	53.92 ± 1.16	47.58 ± 1.62	52.20 ± 1.45	39.43
No ARC Large DE	23.48 ± 1.26	32.30 ± 0.98	30.45 ± 0.48	56.09 ± 1.16	60.32 ± 1.59	52.11 ± 1.45	42.46
Large DE	24.98 ± 1.28	34.87 ± 1.00	32.28 ± 0.48	59.79 ± 1.14	61.26 ± 1.58	51.52 ± 1.45	44.12
Small DE + Large EN	23.83 ± 1.26	30.66 ± 0.97	29.69 ± 0.47	56.37 ± 1.16	60.84 ± 1.58	51.01 ± 1.45	42.07
Small DE + DCLM Filter EN (L)	24.45 ± 1.28	32.96 ± 0.99	29.46 ± 0.47	57.07 ± 1.15	60.11 ± 1.59	49.75 ± 1.45	42.30
Small DE + Large HS EN	23.75 ± 1.26	31.81 ± 0.98	30.32 ± 0.47	57.40 ± 1.15	59.68 ± 1.59	51.52 ± 1.45	42.41
Small DE +Large ARC EN	23.83 ± 1.26	33.05 ± 0.99	29.59 ± 0.47	56.37 ± 1.16	60.11 ± 1.59	51.60 ± 1.45	42.43
Small DE + Large S EN	23.66 ± 1.26	32.21 ± 0.98	28.84 ± 0.47	56.26 ± 1.16	61.58 ± 1.58	52.96 ± 1.45	42.58
Small DE + Large S+I EN	23.22 ± 1.25	33.67 ± 0.99	29.42 ± 0.47	55.93 ± 1.16	61.68 ± 1.58	51.27 ± 1.45	42.53
Small DE + Large v1	23.92 ± 1.27	35.27 ± 1.01	32.79 ± 0.49	59.74 ± 1.14	62.42 ± 1.57	50.68 ± 1.45	44.14
Small DE + Large v2	24.27 ± 1.27	36.19 ± 1.01	32.72 ± 0.48	59.41 ± 1.15	63.89 ± 1.56	52.53 ± 1.45	44.84
Small DE + Large v3	23.83 ± 1.26	36.19 ± 1.01	34.06 ± 0.49	61.04 ± 1.14	63.26 ± 1.56	51.77 ± 1.45	45.03
Small DE + Large RPJv2	24.10 ± 1.27	33.10 ± 0.99	30.06 ± 0.47	55.44 ± 1.16	59.05 ± 1.60	50.08 ± 1.45	41.97
Small DE + Large RFW	25.33 ± 1.29	32.12 ± 0.98	30.56 ± 0.48	57.34 ± 1.15	62.32 ± 1.57	51.77 ± 1.45	43.24
Small DE + Large FWE	25.59 ± 1.29	35.00 ± 1.00	31.19 ± 0.48	56.86 ± 1.16	62.32 ± 1.57	51.52 ± 1.45	43.75

Table 14: Evaluation of 300M parameter model on "General Understanding Tasks" focusing on general reasoning, language understanding, and science knowledge in translated German. Results show the length normalized accuracy for individual datasets and the average over all datasets for all datasets.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WGrande	AVG
Small DE	20.73 ± 1.18	27.53 ± 0.92	25.94 ± 0.44	52.88 ± 1.16	36.80 ± 1.53	47.12 ± 1.40	35.17
No ARC Large DE	21.33 ± 1.20	35.40 ± 0.98	30.88 ± 0.46	59.14 ± 1.15	60.50 ± 1.55	50.67 ± 1.41	42.99
Large DE	22.78 ± 1.23	37.92 ± 1.00	33.06 ± 0.47	62.35 ± 1.13	64.70 ± 1.51	51.14 ± 1.40	45.33
Small DE + Large EN	25.94 ± 1.28	48.15 ± 1.03	48.50 ± 0.50	71.55 ± 1.05	73.10 ± 1.40	52.01 ± 1.40	53.21
Small DE + Large EN DCLM Filter	31.06 ± 1.35	57.24 ± 1.02	49.26 ± 0.50	71.71 ± 1.05	78.40 ± 1.30	54.06 ± 1.40	56.96
Small DE + Large HS EN	28.84 ± 1.32	49.83 ± 1.03	54.83 ± 0.50	75.14 ± 1.01	75.10 ± 1.37	54.85 ± 1.40	56.43
Small DE +Large ARC EN	31.83 ± 1.36	57.24 ± 1.02	48.80 ± 0.50	73.07 ± 1.04	77.00 ± 1.33	52.72 ± 1.40	56.78
Small DE + Large S EN	30.63 ± 1.35	57.28 ± 1.02	39.79 ± 0.49	68.28 ± 1.09	80.40 ± 1.26	53.91 ± 1.40	55.05
Small DE + Large S+I EN	29.10 ± 1.33	55.22 ± 1.02	48.34 ± 0.50	71.71 ± 1.05	78.50 ± 1.30	53.04 ± 1.40	55.98
Small DE + Large v1	20.73 ± 1.18	38.97 ± 1.00	36.09 ± 0.48	63.33 ± 1.12	65.40 ± 1.51	52.01 ± 1.40	46.09
Small DE + Large v2	22.70 ± 1.22	38.97 ± 1.00	36.55 ± 0.48	64.80 ± 1.11	69.30 ± 1.46	51.22 ± 1.40	47.26
Small DE + Large v3	20.65 ± 1.18	40.36 ± 1.01	36.57 ± 0.48	63.66 ± 1.12	70.50 ± 1.44	51.54 ± 1.40	47.21
Small DE + Large RPJv2	26.96 ± 1.30	50.42 ± 1.03	51.10 ± 0.50	70.57 ± 1.06	77.60 ± 1.32	55.64 ± 1.40	55.38
Small DE + Large RFW	27.90 ± 1.31	54.59 ± 1.02	54.91 ± 0.50	73.23 ± 1.03	77.60 ± 1.32	56.35 ± 1.39	57.43
Small DE + Large FWE	38.14 ± 1.42	66.37 ± 0.97	54.88 ± 0.50	72.25 ± 1.04	84.60 ± 1.14	56.04 ± 1.39	62.05

Table 15: Evaluation of 1.3B parameter model on "General Understanding Tasks" focusing on general reasoning, language understanding, and science knowledge in English. Results show the length normalized accuracy for individual datasets and the average over all datasets for all datasets.

Model Name	ARC-C-DE	ARC-E-DE	HS-DE	PIQA-DE	SCIQ-DE	WGrande-DE	AVG-DE
Small DE	22.69 ± 1.24	30.31 ± 0.97	28.99 ± 0.47	54.84 ± 1.16	40.00 ± 1.59	50.42 ± 1.45	37.88
No ARC Large DE	25.07 ± 1.29	36.95 ± 1.02	35.64 ± 0.49	59.74 ± 1.14	64.53 ± 1.55	52.62 ± 1.45	45.76
Large DE	27.44 ± 1.32	38.81 ± 1.03	39.53 ± 0.51	63.17 ± 1.13	67.68 ± 1.52	52.62 ± 1.45	48.21
Small DE + Large EN	25.07 ± 1.29	37.70 ± 1.02	36.18 ± 0.50	59.36 ± 1.15	63.05 ± 1.57	50.51 ± 1.45	45.31
Small DE + Large EN DCLM Filter	26.21 ± 1.30	40.09 ± 1.03	35.86 ± 0.50	59.30 ± 1.15	64.00 ± 1.56	51.94 ± 1.45	46.23
Small DE + Large HS EN	24.89 ± 1.28	36.90 ± 1.02	37.11 ± 0.50	60.77 ± 1.14	63.47 ± 1.56	52.87 ± 1.45	46.00
Small DE +Large ARC EN	26.91 ± 1.32	39.34 ± 1.03	35.77 ± 0.50	58.65 ± 1.15	67.58 ± 1.52	52.11 ± 1.45	46.73
+ Large S EN	27.35 ± 1.32	39.38 ± 1.03	33.23 ± 0.49	58.76 ± 1.15	64.21 ± 1.56	51.18 ± 1.45	45.69
Small DE + Large S+I EN	25.24 ± 1.29	39.38 ± 1.03	35.34 ± 0.49	59.74 ± 1.14	64.21 ± 1.56	52.70 ± 1.45	46.10
Small DE + Large v1	26.12 ± 1.30	40.93 ± 1.03	40.39 ± 0.51	61.48 ± 1.14	67.58 ± 1.52	51.94 ± 1.45	48.07
Small DE + Large v2	25.51 ± 1.29	42.30 ± 1.04	40.75 ± 0.51	62.19 ± 1.13	71.37 ± 1.47	52.45 ± 1.45	49.09
Small DE + Large v3	26.21 ± 1.30	40.84 ± 1.03	43.08 ± 0.51	64.09 ± 1.12	66.74 ± 1.53	51.52 ± 1.45	48.75
Small DE + Large RPJv2	24.71 ± 1.28	37.17 ± 1.02	36.53 ± 0.50	58.65 ± 1.15	65.26 ± 1.55	52.70 ± 1.45	45.84
Small DE + Large RFW	25.42 ± 1.29	38.81 ± 1.03	38.25 ± 0.50	58.38 ± 1.15	64.21 ± 1.56	51.01 ± 1.45	46.01
Small DE + Large FWE	26.91 ± 1.32	42.39 ± 1.04	37.40 ± 0.50	60.45 ± 1.14	65.37 ± 1.54	50.42 ± 1.45	47.16

Table 16: Evaluation of 1.3B parameter model on "General Understanding Tasks" focusing on general reasoning, language understanding, and science knowledge in translated German. Results show the length normalized accuracy for individual datasets and the average over all datasets for all datasets.