Assessing the Role of Data Quality in Training Bilingual Language Models

Skyler Seto, Maartje ter Hoeve*, Maureen de Seyssel*, David Grangier*
Apple
{sseto,m_terhoeve,mdeseyssel,grangier}@apple.com

Abstract

Bilingual and multilingual language models offer a promising path toward scaling NLP systems across diverse languages and users. However, their performance often varies wildly between languages as prior works show that adding more languages can degrade performance for some languages (such as English), while improving others (typically more data constrained languages). In this work, we investigate causes of these inconsistencies by comparing bilingual and monolingual language models. Our analysis reveals that unequal data quality, not just data quantity, is a major driver of performance degradation in bilingual settings. We propose a simple yet effective data filtering strategy to select higher-quality bilingual training data with only high quality English data. Applied to French, German, and Chinese, our approach improves monolingual performance by 2–4% and reduces bilingual model performance gaps to 1%. These results highlight the overlooked importance of data quality in multilingual pretraining and offer a practical recipe for balancing performance.

1 Introduction

Language models (LMs) exhibit exceptional performance on a number of language understanding and knowledge tasks (Brown et al., 2020; Bubeck et al., 2023; OpenAI, 2023). While much of the effort in training language models is focused solely on English, recent bi- or multilingual models also incorporate other languages (De Vries et al., 2019; Martin et al., 2019; Wei et al., 2023a; Faysse et al., 2024; Lample and Conneau, 2019; Xue et al., 2021; Workshop et al., 2022; Yang et al., 2024). In many scenarios, it is beneficial to train a multilingual model as (i) maintaining a separate model for each language can be costly in memory and inference constrained settings, (ii) data relevant to some tasks may only be available in specific languages, and (iii) for many languages, the amount of available high quality data is insufficient for pretraining monolingual language models.

In contrast to work in the vision domain, which shows scaling data (even noisy or from different domains) improves model performance (Sun et al., 2017), prior work in language modeling shows that multilingual models are prone to degradations relative to monolingual models (Chang et al., 2024; Conneau et al., 2020; Xu et al., 2024). Fundamentally, training multilingual models requires learning the structure and semantics of each language. Thus, such models may require training for longer (Conneau et al., 2020; Chang et al., 2024) and on better data to reach the same performance as prior work shows that the sample complexity of learning from multi-distribution data grows with the number of distributions (Haghtalab et al., 2022). Critically, there are two main deficits in prior investigations:

Data quality. While prior work studies how data size impacts performance degradations in multilingual models, they do not study data quality in multilingual models (Chang et al., 2024). Previous papers define data quality according to a few core principles: (i) fluent language (Penedo et al., 2023; Raffel et al., 2020), (ii) long form text (Li et al., 2023; Yang et al., 2024), and (iii) informative text with educational content and textbook format (Li et al., 2024; Penedo et al., 2024a). We discuss data selection in more detail in Section 2.

Data quality has already been shown to be an important factor in training high performing English language models (Li et al., 2024, 2023; Maini et al., 2024). Because of this, there is growing interest in model-based filters for curating high quality data in monolingual settings (Li et al., 2024; Messmer et al., 2025; Penedo et al., 2024a). There are many other languages for which training a language model is practical as a reasonably large amount

*Equal contribution

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22694–22720
November 4-9, 2025

of unfiltered data is available (Weber et al., 2024; Penedo et al., 2024b). However, the importance of data quality for training multilingual models and data quality filtering in multilingual settings has received little attention. There are several challenges that can arise from filtering high quality data for multilingual pretraining as (i) quality filters may work differently across languages, (ii) the density of high quality data and filtered topics may vary, and (iii) the impact of quality filtering may have a small impact across languages in multilingual settings.

Model and data size Prior works that study gaps in multilingual model performance typically aim to study performance gaps from training on a large number of languages (typically on the order of hundreds), small data per language, and smaller encoder-style architectures (Conneau et al., 2020; Chang et al., 2023; Xu et al., 2024). In this setting, they refer to the gap as the curse of multilinguality. However, studying gaps in performance in these settings greatly impacts the evaluations that are feasible, makes it difficult to control training data at scale, and does not control for the fact that training a multilingual model is simply a more challenging task, and may require longer training time or higher capacity to achieve the same performance with the same amount of data.

This work focuses on exploring how these challenges underlie gaps in multilingual performance. We conduct experiments on data quality (measured by information/knowledge) and language in French and Chinese, where we control both through training on translated data. We also provide a recipe for obtaining high quality data in other languages that improves bilingual model performance in three languages: French, German, and Chinese[1]. Collectively, our main contributions show that (i) High quality data filtering in multiple languages without access to native high quality data improves performance in the target language, and reduces gaps in monolingual and bilingual performance. (ii) Data quality plays an important role in the performance of bilingual language models (rather than only the language or data size). (iii) High quality English data alone is insufficient for training high performing multilingual language modeling in some tasks.

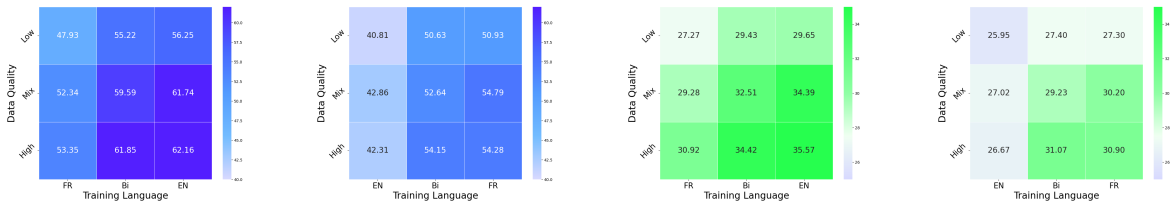
[1]We select these languages for their use in prior work, distance from English, amount of data available, and availability of evaluation benchmarks.

2 Related Work

Multilingual Language Models Large scale multilingual language models are of two main types: (i) They can be trained on a large corpus of multilingual data such as mC4 (Xue et al., 2021), CCNet (Wenzek et al., 2020), or FineWeb2 (Penedo et al., 2024b) typically covering in the order of 100 languages. This includes models such as mBert (Devlin et al., 2019) , XLM (Conneau et al., 2020), mT5 (Xue et al., 2021), Bloom (Workshop et al., 2022), etc. (ii) Bilingual language models such as 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), which are typically small, but can be large in the case of Chinese where an abundance of high quality data is present (Yu et al., 2025). Other models such as the Llama family (Touvron et al., 2023), Mistral (Jiang et al., 2023), Gemini (Team et al., 2023), Palm2 (Anil et al., 2023), and GPT (OpenAI, 2023) have been shown to have multilingual capabilities, however a majority of their data is English (Xu et al., 2024).

Multilingual Data Curated datasets are essential to training language models. Early multilingual datasets include CCNet (Conneau et al., 2020), mC4 (Xue et al., 2021), and CulturaX (Nguyen et al., 2024) all support over 100 languages, though the largest sources of data are English and even other high resource languages such as German, French, Chinese, and Korean contain 10-50× less data. Other datasets such as Redpajamav2 contain over 2.5T tokens, but are limited to only a few Indo-European languages (Weber et al., 2024), and still a factor of 7-10× less than English. Recently the FineWeb2 dataset was crafted for many languages with the same heuristic filters as the original FineWeb supporting many languages with data for pretraining (Penedo et al., 2024b). Still, there is only a handful of high quality datasets large enough for training language models in select languages like Chinese (Yu et al., 2025), French, German, and Spanish (Penedo et al., 2024b; Messmer et al., 2025).

Data Selection High quality data selection remains an important area of research in training language models. Early research on data selection was based on heuristics including GPT-2 (Radford et al., 2019), Gopher (Rae et al., 2021), C4 (Raffel et al., 2020), and RefinedWeb (Penedo et al.,



(a) English Core (b) French Core (c) English MMLU (d) French MMLU

Figure 1: Performance with varying data quality and language. Models are trained on combinations of mC4 (low) and FineWebEDU (high) in native English (EN) and translated to French (FR). Models are trained for 200K steps and evaluated on Core (avg over six common-sense reasoning tasks) and MMLU.

2023). Recent works examine model based filters for labeling general high quality data (Sachdeva et al., 2024; Li et al., 2024), or textbook quality data (Penedo et al., 2024a). While a majority of this work focuses on English data only, a few works have examined filtering in other languages, such as by perplexity (Conneau et al., 2020) or filtering using models trained on high quality specialized data in select languages (Messmer et al., 2025). Other forms of data selection include reweighting data (Grangier et al., 2024a; Fan et al., 2023; Xie et al., 2024) and are shown to have varying degrees of success when applied in bilingual settings with data constraints, but only for English (Seto et al., 2024).

3 Data Quality and Multilinguality

We conduct four types of experiments to demonstrate that the performance gap between a bilingual language model and monolingual models is largely due to the data quality and number of training steps - e.g., multilingual models require better training and for longer. We start with demonstrating that there are performance gaps when not controlling for data quality (Section 3.2). We then show that training on a translated pretraining corpus in both languages, thereby controlling data quality, yields no gap between monolingual and bilingual performance (Section 3.3). Next, we find that at smaller number of training steps, there is a gap between multilingual and monolingual models, and models learn faster with higher quality data (Section 3.4). For the experiments controlling data quality, we show results for English and French translated data here, and refer to Section E for English and German experiments. Finally, we show that quality also depends on the information available in the data, and that high quality English data with translations alone is insufficient for training high performing bilingual models (Section 3.5). These experiments are done with Chinese and English given the availability of high quality Chinese data for trainng and downstream evaluations similar to MMLU.

3.1 General Model Details

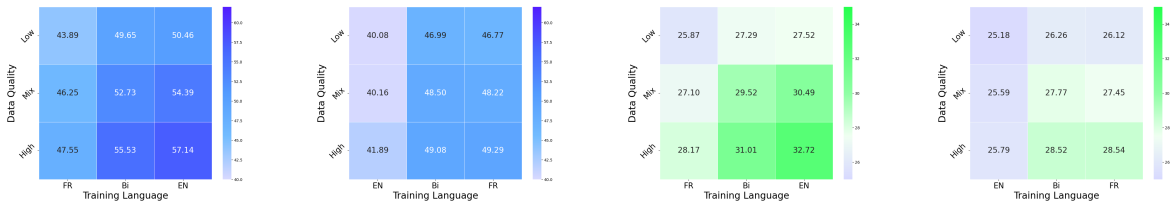
We train decoder-only transformer models (Vaswani et al., 2017) with 1.3B non-embedding parameters. Models use the PolyLM tokenizer (Wei et al., 2023b), with a total vocabulary size of 256K tokens using BPE to allow for using the same tokenizer across all experiments. Models are trained for 200K steps with batch size 1024 and context length 1024 unless otherwise stated. Hyperparameters and model details are in Appendix A. This model size is chosen as it provides reasonable (above random) performance on several benchmark QA tasks, and is commonly used for benchmarking and ablating pretraining of language models (Penedo et al., 2023, 2024a).

3.2 Model Performance without Controlled Data Quality

Methodology: We start with a setup in which a bilingual language model is trained on an equal proportion of data from mC4 in French (FR) and English (EN) totaling 100K steps each.

Model	Core EN	Core FR	MMLU EN	MMLU FR
EN	**56.3**	40.8	**29.7**	26.0
FR	45.9	49.0	27.0	27.2
BI	53.5	**49.6**	29.3	**27.4**

Table 1: Zero shot accuracy for general understanding and specialized knowledge tasks for monolingual English (EN), French (FR), and bilingual (BI) models.



(a) English Core (b) French Core (c) English MMLU (d) French MMLU

Figure 2: Performance with varying data quality and language. Models are trained on combinations of mC4 (low) and FineWebEDU (high) in native English (EN) and translated to French (FR). Models are trained for 30K steps and evaluated on Core and MMLU.

Findings: Table 1 shows performance on MMLU and Core[2] benchmarks. Our findings match those in prior works including (Conneau et al., 2020; Chang et al., 2024) where we see a 3% drop in English and an increase in French of 0.8% compared to the bilingual model for Core tasks. For MMLU, the difference is smaller but the trends remain similar. The bilingual model has the same ratio of data and sufficient data for training ($\sim 7\times$ Chinchilla).

3.3 Model Performance Varying Data Quality

Methodology: To demonstrate that multilingual performance depends on data quality, we now control the data quality and languages in the model. We follow the same setup as above and vary both the data quality and the language. For our experiments, we use the datasets: FineWebEDU, and mC4 EN. These datasets are chosen as they have varying quality according to the DCLM classifier[3] (mean quality scores 0.023 vs. 0.1127 respectively), and the FineWebEDU dataset has much higher performance on downstream benchmarks.

We translate mC4 into French using a proprietary translation system following (Seto et al., 2024), and use TransWebEDU translations for FineWebEDU (Wang et al., 2025). We consider a variety of scenarios where the quality can vary, or the language can vary, and measure the performance on both English and French.

Findings: Figure 1 shows the performance difference when varying quality (y-axis) and language (x-axis) for two sets of zero-shot evaluations: Core and MMLU. for all evaluations we use the continuation version of the task. We denote training with the mC4 dataset as low quality, and FineWebEDU as high quality. When examining the plots, we see that the bottom right square corresponding to a monolingual model[4] trained on high quality in the targeted language has the highest performance.

This is closely followed by models which individually vary the language but keep high quality [middle bottom square, e.g., (bi, high)] or mix quality but keep the same language [right middle square, e.g., (EN, mix)], which are all within 1%. However, mixed quality and mixed language taken together [middle square, e.g., (bi, mix)] exhibits an average 2% drop in English performance by comparison, compared to each on their own[5], and all squares with low quality (top row) exhibit a much larger drop than bilingual models with mixed or high quality (bottom four squares).

3.4 Model Performance with Fewer Steps

Methodology: Next, we show that training for fewer steps yields a gap between bilingual and native model performance. Specifically, the experimental setup is the same as above, but we examine training after 30K steps equating to roughly Chinchilla scaling for a 1.3B model.

Findings: Results are shown in Figure 2. At this scale, we see that both the bilingual high quality (middle bottom) and mixed quality native monolingual (right middle) models have 2-2.5% lower performance than the monolingual high quality unlike prior results at 200K steps for English evaluations. Similarly low quality results (top row) drop below

[2]Average over six general knowledge and common-sense reasoning tasks: ARC-easy, ARC-challenge, SciQ, PIQA, HellaSwag, Winogrande

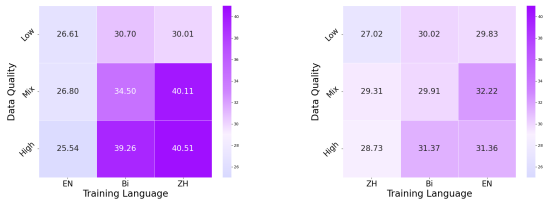
[3]A fasttext classifier aimed at distinguishing high quality data according to samples found in OpenHermes and highly upvoted ELI5 posts - https://huggingface.co/mlfoundations/fasttext-oh-eli5.

[4]We flip the x-axis order depending on the evaluation task such that the bottom right is always the high quality monolingual model for consistent comparison.

[5]Note that the middle square represents the average of two models: as both languages could have the high quality data source

bilingual again indicating data quality has a large role in training.

3.5 Model Performance with High Quality Data in Multiple Languages



(a) CMMLU ZH (b) CMMLU EN

Figure 3: Performance with varying data quality and language. Models are trained on combinations of Chinese FineWebEDU (high) and FineWebEDU (low) in native English (EN) and translated to Chinese (ZH). Models are evaluated on CMMLU.

Methodology: This section discusses the impact of training models on two datasets in different languages that are collected by similar filtering (classification of textbook and science knowledge). We repeat the same training recipe with two highly curated datasets in different languages: Chinese FineWebEDU and FineWebEDU (which is in English). For this experiment, we refer to the FineWebEDU dataset as "low" and Chinese FineWebEDU as "high". Here we refer to the drop in quality as (iii) informative text, where FineWebEDU does not cover topics relevant to CMMLU. This is in contrast to previous sections, for which all definitions of quality drop. These datasets are both curated in the same way, and considered high quality in their respective languages, however may contain culturally different information. We translate both datasets into the other language using the same proprietary translation system as in (Seto et al., 2024).

Findings: Figure 3 shows the performance difference for monolingual and bilingual models trained in Chinese and English. We find that bilingual high quality models trained on native Chinese data (middle bottom square) drop in performance by $\sim 1\%$, but still perform better than translated English data to Chinese (right top square). This drops slightly from the mixed monolingual models (right middle square) indicating there may be some effect from further languages. Nonetheless, our findings show that translated data from English alone may not be sufficient for high quality, and building high quality datasets through filtering with the same mechanisms as in English can help yield bilingual models that also perform well in non-translated tasks.

4 Language-Agnostic Data Filtering

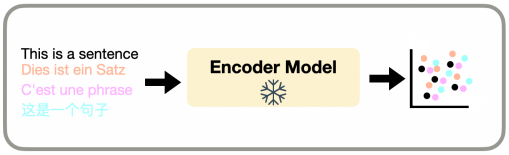
Section 3 shows that bilingual models may perform as well as monolingual models when the data used to train the models has sufficient information for the target downstream tasks, and is of comparable quality. However, for a large set of languages, high quality data does not exist. Prior works have shown that learnable models as quality filters lead to improved performance in downstream tasks for English (Li et al., 2024; Grangier et al., 2024a; Penedo et al., 2024b).

To learn a model-based filter for selecting high quality data, we assume access to a small set of high quality data $\mathcal{D}^h = \{(x, y)|x \in \mathbb{R}^d, y \in \{0, 1\}\}$, where x is some representation of a document, and y is a binary label indicating the quality of the sample. A binary classifier ϕ is trained on $\mathcal{D}^h$ to estimate the probability that a document from the general pretraining data $\mathcal{D}^g$ is high quality. The high quality pretraining set D^{hq} is then selected according to the classification rule

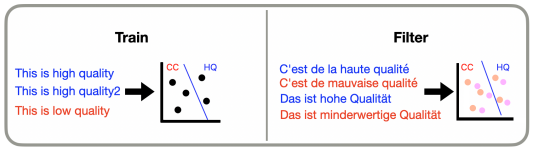
$$\mathcal{D}^{hq} = \{x \in \mathcal{D}^g | \phi(x) > \tau\}, \quad (1)$$

for some predefined threshold τ. Unlike prior works which assume training the classifier and selecting general high quality data (Li et al., 2024), or specialized data relevant to downstream tasks (Grangier et al., 2024a) in English, we assume the high quality data sample is available only in English, but will be used to select data in other languages. Concurrent to our work, (Messmer et al., 2025) show that a classifier approach in (Grangier et al., 2024a) can be applied to languages other than English. They follow a similar setup using specialized datasets, such as translated MMLU (Singh et al., 2024), and Include (Romanou et al., 2024), for data selection. As such it is still unclear whether a universal language embedding with high quality seed data available only in English can be used to train a language agnostic filter, and whether monolingual filtering improves bilingual model training. We provide preliminary experiments indicating similar distributions of English and translated French data in Appendix D.

In our experiments, we parameterize ϕ as a logistic regression, and use a lightweight Sentence-



(a) Multilingual Representations



(b) EN Train and Multilingual Filter

Figure 4: (a) **Multilingual Language Representations**: Build a universal sentence embedding that maps multilingual data to the same embedding space. (b) **English High Quality Training and Multilingual Data Filtering**: Classifier is trained on the embeddings of a small amount of high quality data available in English. The classifier is then used to filter data in all supported languages.

BERT (SBERT) multilingual model[6] for extracting features (Reimers and Gurevych, 2019). For $\mathcal{D}^h$, we use the same English data used for training the DCLM classifier, which compares data from RefinedWeb (low quality), and OpenHermes 2.5 or ELI5 (high quality). In our ablations, we also explore using the annotations for FineWebEDU which scores documents from 1 to 5 based on educational content. We do binary classification using a score above 2 as high quality.

The value of τ is selected to ensure enough data for pretraining, and is in the order of 10% of the data following (Li et al., 2024). In this work, we train a classifier as we have limited data within each cluster for training a 1.3B model at the desired scale, and would repeat data significantly if we instead do data selection for specialized data.

5 Experiments

5.1 Experimental Setup Details

We use the same 1.3B parameter models trained for 200K steps as in Section 3 and train with FineWeb2 and Redpajama2 datasets for filtering. Additional details on exact pools of data are available in Appendix B and different model sizes in Section K.1. We give our main experiments and ablations on English-French bilingual pretraining and include German and Chinese filtering in Section J. Additional motivation for language selection is in Appendix B. Individual task accuracy in Appendix L.

5.2 Analysis of Model Based Filter Selection

Precisely measuring the performance of a data quality classifier is difficult as there is no ground truth. Our main constraint is that the classifier scoring is tied to performance on downstream tasks. Before we discuss performance in other languages, we first show that the acceptance rate of the classifier matches commonly accepted notions of high quality datasets.

In particular, we consider various English datasets and measure the amount of high quality data using a fixed threshold of 0.14094 which corresponds to filtering down to 10% of the data in C4. Our findings shown in Table 2 for the amount of high quality in each dataset match the performance of 1.3B models trained in ablations for FineWebEDU, and correspond to the amount of filtering in each dataset (Penedo et al., 2024a).

Source	% High Quality Data
C4	10.00%
RPJv2	12.93%
RFW	15.40%
Wiki	35.01%
FineWebEDU	47.94%

Table 2: Percent of data which is considered high quality in different English datasets

We also observe, in the case of Chinese, that our English-trained classifier correctly identifies better quality corpora. Using a threshold of 0.2751, we find that 3.38% of mC4, 10% of FW2, and 33.53% of ChineseFineWebEDU is considered high quality.

Additionally, it is important to have a good understanding of the factors that contribute to a sample being high quality for language model training. We evaluate two commonly used linguistic measures: the mean cosine similarity between pairs of contiguous sentences in the same document following (Barzilay and Lapata, 2008), and the Flesch reading ease as a measure of how easy-to-read the text is (Kincaid et al., 1975).

We compute the Pearson correlation between the filter score and each measure over 1000 samples. For coherence, we have mean for low quality is 44.78, the mean for high quality is 48.14,

[6]The model, `paraphrase-multilingual-MiniLM-L12-v2`, is at https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2.

corr=0.1421, and p-val=$6.48e-06$, and for reading ease, the mean for low quality is 0.346, the mean for high quality is 0.365, corr=0.1464, and p-val=$3.32e-06$ meaning that as filter score increases (more high quality) our linguistic measures suggest greater coherency, and higher reading ease.

5.3 Monolingual Data Filtering Performance

Methodology: We show that increasing quality according to our model-based filter leads to an increase in downstream task performance in monolingual settings. For our experiments, we train a 1.3B parameter model for 30K steps, and evaluate the zero-shot accuracy on the Core set of tasks. We compare the SBERT filtering classifier with training a model on raw data from RedPajama2, and FineWebEDU in the respective language, as well as filtering RedPajama2, and using a fasttext classifier for filtering trained on the DCLM classifier training data data translated in French. We use two translation systems to show the effect of translation: a cheap proprietary CPU translation system, and the Mistral-7B model following (Wang et al., 2025). We show results for full comparisons for 30K steps (Figure 5), and for 200K steps in Appendix F.1.

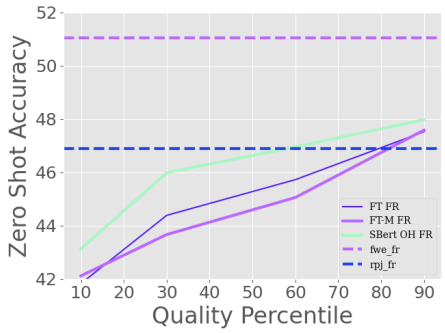


Figure 5: Quality vs. accuracy on Core tasks for filtered RedPajama2 (SBert OH FR) compared with filtering using a FastText classifier (FT FR and FT-M FR), TransWebEDU (fwe_fr) and base RedPajama2 (rpj_fr) in French after 30K steps.

Findings: As we increase the percentile of data quality, the accuracy increases leading to a 6% increase between the lowest and highest quality. Second, filtering with SBERT outperforms training without filtering for RedPajama2, and outperforms a fasttext classifier trained on translated data as in (Li et al., 2024), even with a high quality translation system such as Mistral-7B. Training on the high quality filtered data falls short of the TransWebEDU data, however this is expected to be an upper bound since the evaluations are also translated benchmarks, for which the FineWebEDU data is more highly curated than RedPajama2.

5.4 Filtering from Already Curated Data

Methodology: Next, we show that our method also selects high quality data in more highly curated datasets. We run the filtering on the FineWeb2 French dataset, which has additional heuristic filtering as originally done for the FineWeb2 English data (Penedo et al., 2024b), and compare with RedPajama2 French data. We select the top 10% of data for training for both datasets, noting that the amount of data available in FineWeb2 could be around 10% of the amount of data in RedPajama2 as estimated by the number of words in the corpus..

Findings: Table 4 shows results for monolingual French models. Our results indicate that performance increases even with the smaller FineWeb2 (FW2) dataset and repeated epochs of training. We find that many of the heuristic filters and text extraction also lead to better performance as the base FineWeb2 improves on even the 90% filter over the Redpajama2 (RPJ2) data. Finally, we note that the performance of the filtered FineWeb2 data matches that of TransWebEDU indicating similar performance to highly curated translated English data on translated evaluations. Experiments for other percentiles on FineWeb2 are in Appendix F.

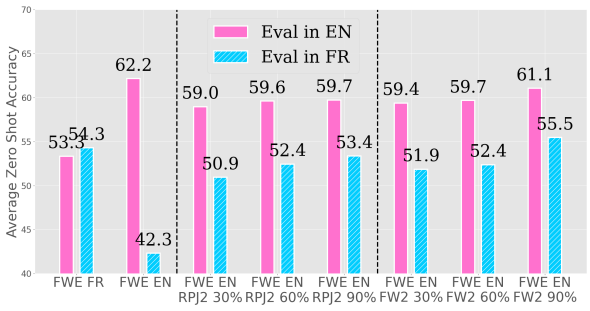


Figure 6: Bilingual vs. Monolingual performance on Core English (EN) and French (FR) benchmarks with Filtering in French. All models use the TransWebEDU English (FWE EN) data while varying the French data.

5.5 Filtering for Bilingual Models

Methodology Now, we show in addition to monolingual performance gains, our data selection method diminishes the gap in bilingual models matching the lack of performance gap we see in Section 3 with native data. For our experiments we compare models trained on data at different quality

Model	Core EN	MMLU EN	Core	MMLU	FB-MC	Regional	NLI	AVG
RPJ2 1.3B	59.74	33.25	52.16	29.35	60.24	31.99	39.72	42.69
FW2 1.3B	60.26	33.44	53.65	29.57	60.24	31.5	39.78	42.95
TWE 1.3B	61.85	34.43	54.15	31.07	54.48	29.55	42.14	42.28
TranswebLLM	59.61	34.26	55.05	30.44	53.00	29.64	40.39	41.70
CroissantLM	56.67	30.59	52.58	28.63	58.89	30.17	41.33	42.32
Bloom 1.1B	51.73	29.10	48.70	27.23	55.16	28.91	40.11	40.02
Qwen2.5 1.5B	69.45	41.14	**55.65**	**32.49**	59.14	31.55	43.55	**44.50**
EuroLLM 1.7B	63.05	35.39	54.94	30.20	60.34	32.54	40.42	43.69
RPJ2 90% 1.3B	59.73	33.56	53.37	29.70	59.81	**32.91**	**43.59**	43.88
FW2 90% 1.3B	61.07	33.54	55.47	30.16	**61.67**	32.82	40.52	44.13

Table 3: Comparison of different bilingual models in French and English compared with other public multilingual models of similar sizes on French evaluation tasks.

Model	RPJ2	FW2	TWE
Base	48.38	51.61	54.28
90% Filter	50.42	54.17	–

Table 4: Performance on Core French benchmarks comparing monolingual models with and without filtering across RedPajama2 (RPJ2), FineWeb2 (FW2), and TransWebEDU French (TWE) datasets.

percentiles at the 30th, 60th, and 90th percentile, and take a total of roughly 10% of the data. We evaluate on the Core evaluations and report results in Figure 6 comparing monolingual TransWebEDU performance with bilingual models.

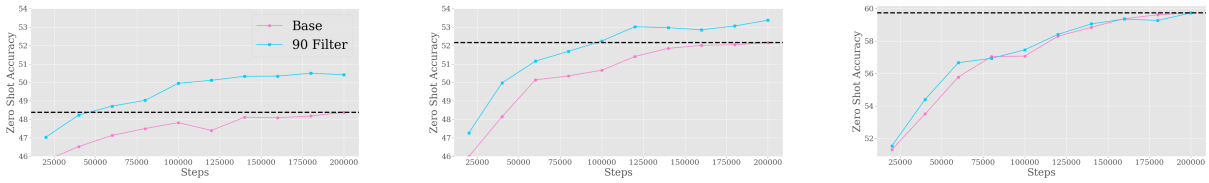
Findings: As we increase data quality, both the EN and FR performance increase. 90% filtering achieves the strongest performance. We note that this is true even over higher filtering where we did train a model at 95% filtering for RedPajama2, but observed over-filtering on the data as training for the same number of steps requires repeating the data which leads to performance leveling out.

We further note that while French evaluations improve consistently over base RedPajama2, and are close to TransWebEDU FR, the English performance is worse and consistent with the base comparison from 60% indicating that the FineWebEDU corpus is still higher quality as it has some additional filtering over RedPajama2. For evaluations filtering FineWeb2, we note that performance in English is within 1%, and performance in French is better than using TransWebEDU. We conclude that improving data quality using our filtering mechanism leads to performance improvements also in English over low quality data.

5.6 Comparison with Bi- and Multilingual Models

Methodology We show that our data selection process in Section 4 can be used to select high quality data consistent with other public bilingual and multilingual models. We study performance for 1.3B parameter models and similar sizes across a range of tasks in English and French as (i) there are a number of available evaluations that are both translated and native, and (ii) there are other bilingual or multilingual models in French for comparison. All of our models are trained with FineWebEDU English as the English data source for 200K steps, which is up to 15× fewer than other models. Additional details on the evaluation sets are provided in Appendix B. For model comparisons, we include strong bilingual models like CroissantLLM (Faysse et al., 2024), models trained on Indo-European languages like TransWebEDU EN-FR, TransWebLLM (Wang et al., 2025), and EuroLLM (Martins et al., 2024), and multilingual models like Bloom (Workshop et al., 2022), and Qwen2.5 (Yang et al., 2024), all of which achieve strong results on multilingual benchmarks.

Findings: Our filtering leads to better zero-shot performance over public bilingual models such as Croissant LLM (1.7%), and competitive performance (up to 4% increase) to multilingual models trained for much longer highlighting the benefits of training a bilingual model on high quality data. Our models attain better performance in French than all models except for Qwen2.5 1.5B which



(a) Monolingual FR (b) Bilingual FR (c) Bilingual EN

Figure 7: Performance at intermediate checkpoints during training for 1.3B models for Core EN and FR benchmarks.

has overall 0.4% improvement while being trained for much longer and on combinations of different data. However, on English data, the Qwen and EuroLLM performance exceed our models.

5.7 Data Scaling

Methodology: Our main results train models for 200K steps amounting to 200B tokens as Section 3 shows benefits from both training for longer and on higher quality toward diminishing bilingual model gap in performance. This amount, although below proprietary models at similar scales such as (Touvron et al., 2023; Yang et al., 2024), is above the recommended training data size according to Chinchilla (Rae et al., 2021) ($\sim$ 7x chinchilla). Li et al. (Li et al., 2024) filter to a small amount of data from a much larger pool (although they filter at the same 10% rate) and train only at 1-2$\times$ Chinchilla scale thus having a much smaller ratio of tokens used. We now show that performance improvements hold at intermediate training steps.

Findings: We show results in Figure 7 for monolingual and bilingual models. Both monolingual and bilingual models evaluated on Core French benchmarks show consistent performance improvements at all stages of training. For monolingual models, performance from filtering leads to around $5\times$ efficiency as the model attains the same performance at only 40K steps. For bilingual models, we note that the higher quality English data reduces that gap consistent with findings on better auxiliary data from (Seto et al., 2024), however we still observe around $2\times$ speedup in training. Finally, for English evaluations, we observe improvements only early in training consistent with bilingual gaps earlier. The gap in English performance diminishes after around 60K steps ($\sim$ 2x Chinchilla).

5.8 Multilingual Model Training

We investigate improvements for multilingual models. We train a 1.3B parameter model for 400K steps with the same ratio per language using the SentenceBert filtered FineWeb2 corpus for Chinese, French, and German, and FineWebEDU as the English corpus. Results are provided in Table 5. We attain similar performance to the bilingual models with filtering. These results are consistent with the bilingual results in Sections 3 and 5.

Model	Core EN	Core FR	Core DE	Core ZH
FWE EN	63.58	-	-	-
FW2 FR 90% + FWE EN	61.07	55.47	-	-
FW2 DE 90% + FWE EN	60.85	-	53.90	-
FW2 ZH 90% + FWE EN	58.67	-	-	53.97
FW2 Multi Base + FWE EN	57.93	51.25	51.00	51.86
FW2 Multi 90% + FWE EN	59.77	54.96	53.86	54.31

Table 5: Comparison of filtering for bilingual and multilingual models for Core benchmarks in the non-english languages. Models are trained with roughly the same amount of data from each language filtered from FineWeb2 for the same number of steps. For English data, we use FineWebEDU.

6 Conclusion

Training a multilingual language model that performs as well as monolingual models is important for building language models that can work for everyone, and facilitate compute efficiency in memory constrained settings where keeping many monolingual models may be infeasible. However it is also more challenging as it necessitates learning multiple distributions of data. This work provides a simple recipe for selecting high quality data, and demonstrates capability of selecting high quality data in other languages with only high quality English data. Selecting high quality data with our recipe reduces gaps between monolingual and bilingual models to less than 1%, and improves monolingual performance. Our work takes a step towards pretraining language models in languages with limited high quality data, and can help more research into closing the gap between multilingual and English-centric language models.

7 Limitations

This section lists limitations of our work.

Evaluation data. Our evaluations languages other than English rely on translated evaluation sets. Not only does this introduce potential translation mistakes (for example for math or certain scientific terms), the resulting evaluation set also contains cultural biases as has been noted in datasets such as MMLU (Singh et al., 2024). 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. We focus on French as there are many native language benchmarks for which models perform well.

Languages included. The focus of this work is on training bilingual language models. We note that there are several languages for which training a bilingual or multilingual language model is now practical given the size and available training data. However, our goal is in building high quality datasets and showing gaps in performance from lack of data quality control which require filtering from a large pool of data. Training even a 1.3B parameter model at our scale requires 200B+ tokens of data and filtering down to 10% of the data leaves only languages with over 2T tokens (for one repetition training), for which there are few. As we are constrained by having a large pool of tokens with relatively little filtering (Redpajama2), a more highly curated set of tokens (FineWeb2), and native evaluations, for our ablations and studies, this left only a few languages: French, German, and Chinese. We study French in the main text as it satisfies all conditions, and is relatively close to English indicating potential transferability as shown in (Seto et al., 2024). We further note that another constraint is the number of languages in our SentenceBert embeddings. The multilingual SentenceBert model used in this work supports 50+ languages[7]. While this already covers many more languages than high performing language models in those languages, there are methods for adding languages to a multilingual embedding via knowledge distillation (Reimers and Gurevych, 2020).

[7]A full list of language codes: ar, bg, ca, cs, da, de, el, en, es, et, fa, fi, fr, fr-ca, gl, gu, he, hi, hr, hu, hy, id, it, ja, ka, ko, ku, lt, lv, mk, mn, mr, ms, my, nb, nl, pl, pt, pt-br, ro, ru, sk, sl, sq, sr, sv, th, tr, uk, ur, vi, zh-cn, zh-tw.

References

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

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*.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

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, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10822–10837.

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*.

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

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*.

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

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

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

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*.

Nika Haghtalab, Michael Jordan, and Eric Zhao. 2022. On-demand sampling: Learning optimally from multiple distributions. *Advances in Neural Information Processing Systems*, 35:406–419.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

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*.

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*.

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

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*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2024. Eurollm: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*.

Bettina Messmer, Vinko Sabolčec, and Martin Jaggi. 2025. Enhancing multilingual llm pretraining with model-based data selection. *arXiv preprint arXiv:2502.10361*.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao.

2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

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.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2024. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237.

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. 2024a. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

Guilherme Penedo, Hynek Kydlıcek, Vinko Sabolcec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024b. Fineweb2: A sparkling update with 1000s of languages, december 2024b. *URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-2*.

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*.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495.

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 (EMNLP)*, pages 4512–4525.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, et al. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.

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

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.

Israfel Salazar, Manuel Fernández Burda, Shayekh Bin Islam, Arshia Soltani Moakhar, Shivalika Singh, Fabian Farestam, Angelika Romanou, Danylo Boiko, Dipika Khullar, Mike Zhang, et al. 2025. Kaleidoscope: In-language exams for massively multilingual vision evaluation. *arXiv preprint arXiv:2504.07072*.

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

Skyler Seto, Maartje ter Hoeve, Richard He Bai, Natalie Schluter, and David Grangier. 2024. Training bilingual lms with data constraints in the targeted language. *arXiv preprint arXiv:2411.12986*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. 2024.

Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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*.

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.

Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, David Adelani, Yihong Chen, Raphael Tang, and Pontus Stenetorp. 2025. Multilingual language model pretraining using machine-translated data. *arXiv preprint arXiv:2502.13252*.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. *arXiv preprint arXiv:2010.03017*.

Maurice Weber, Dan Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, et al. 2024. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems*, 37:116462–116492.

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*.

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

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

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.

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*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yijiong Yu, Ziyun Dai, Zekun Wang, Wei Wang, Ran Chen, and Ji Pei. 2025. Opencsg chinese corpus: A series of high-quality chinese datasets for llm training. *Preprint*, arXiv:2501.08197.

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 Additional Training Details

The small model is a 350M non-embedding parameter model consisting of 24 layers, 16 attention heads, and a hidden dimension size of 1024. The 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 2.7B parameter model consists of 32 layers with 2560 hidden dimension and 32 attention heads.

The baseline models are trained using NVIDIA's Megatron-LM[8] repository for pretraining language models. All models are trained for a total of 200K steps with a batch size of 1024. The 2.7B models are trained with context of 2048 and other models are trained with 1024.

Models are trained using a maximum learning rate of 0.0003 for the 350M model, 0.0002 for the 1.3B model, and 0.00016 for the 2.7B models with 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 total training time for 1.3B models on roughly 200B tokens is around 2000 GPUh on Nvidia H100 GPUs. For 350M models, the total training time is around 1200 hours. For a 2.7B model trained on roughly 400B tokens, the total time is around 9000 GPUh on Nvidia H100.

B Dataset Details

B.1 Training Datasets

We consider several datasets in this work and primarily focus on FineWeb2, and RedPajamav2 for pretraining. We choose these datasets as there exist a sufficiently large amount of data in multiple languages. For experiments filtering high quality data, we focus on Redpajamav2 as there are up to 3T tokens of data in these datasets compared with mC4 ($\sim$ 300B), and FineWeb2 ($\sim$ 206B words) for French, and the data is native (non-translated). We also experiment with TransWebEDU (Wang et al., 2025) for comparison to training on translated high quality data, and with filtering from FineWeb2, an already filtered but smaller pool of data. We primarily focus on English-French bilingual pretraining in this section as we have both larger amounts of data for pretraining in publicly available corpora such as RedPajama2 and FineWeb2, have native evaluation sets, and the language is relatively close to English. We additionally present results with the high quality filter in German and Chinese in Section J. We choose German as there is also a large amount of high quality data, its closeness to English, and it is commonly used in other works (Seto et al., 2024). However, we note that there is only a small amount of native evaluation data such as Include (Romanou et al., 2024) and Kaleidoscope (Salazar et al., 2025) for which there are only a few hundred samples. We also evaluate on Chinese to test a language further from English with a large amount of publicly available data, and native evaluations. We provide a brief description of each dataset below as well as the token counts for the approximate number of tokens used used for each dataset in training[9].

- **mC4**: We use the multilingual Colossal Clean Crawled Corpus (mC4), a curated text dataset comprising over 6.3T tokens for experiments in Section 3. 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 a low quality dataset as it is relatively little filtering. For our experiments we use the first $\sim$ 520 files for translation and otherwise train on one epoch or two epoch of data from this subset (Xue et al., 2021).

- **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., 2024a). We use the subset known as TransWebEDU, which is a subset of around 75B tokens used in (Wang et al., 2025). We also use the machine translated German version and translate using a proprietary translation system into Chinese in Section 3. We use all files from this dataset given the already smaller size.

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

[9] Note that for training with the Megatron library, we tokenize batches of parquet or jsonl files (referred to as a dataset in Megatron-LM), and use each dataset with equal weight. This means that if some files or documents have fewer tokens, they might repeat at a higher rate than other sets of files. While we do see more repetitions for a few subsets, this is relatively small for overall training, and for training, we still repeat data for only a few epochs less than would incur a gap in performance to single epoch training following (Muennighoff et al., 2024).

- **ChineseFineWeb-EDU**: 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. We use the first 600 files in total for our experiments (Yu et al., 2025).

- **RedPajama2**: A pretraining corpus with light filtering consisting of 30T tokens: 20T tokens of English text, and $\sim$ 3T for German and French. We focus on the French and German portions of the dataset only. We randomly shuffle all subsets of the data and train using a random shuffled subset of both the head and middle portions (Weber et al., 2024).

- **FineWeb2**: Data sourced in a similar way as FineWeb but for many languages. We use French, German, and Chinese subsets. The French data has 113 parquet files, German has 122, and Chinese has 185 parquet files. Given the size of the datasets, data is repeated for multiple epochs though still under 10 epochs to not yield worse performance than training on new data following (Muennighoff et al., 2024).

Dataset	Tokens (B)			
	EN	FR	DE	ZH
mC4 EN	125	76	75	–
RPJ2 (base)	–	310	297	–
RPJ2 (90%)	–	260	248	–
FineWeb2	–	270	260	282
FineWeb2 (90%)	–	34	28	30
TransWebEDU	54	62	55	45
ChineseFineWeb	192	–	–	195

Table 6: Upper bound on the approximate number of tokens by language used in training in this work for training datasets used in this work.

B.2 SentenceBert Filter Scores

We report the SentenceBert filter scores corresponding to different percentiles for all datasets we filter. Filter scores are primarily estimated using only the first file, however we also compare this with filter scores from 100 randomly selected files and find that they are similar. We report the scores used in Tabl 7. Note that for some datasets, we only conduct experiments using the 90th percentile.

B.3 Zero Shot Evaluations

B.3.1 Core Benchmarks

- **SciQ [Core]**: A dataset of science exam questions for evaluating the ability of NLP models in understanding and reasoning within the science domain (Welbl et al., 2017).

- **ARC Challenge (ARC-C) [Core]**: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) [Core]**: 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.

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

- **PIQA [Core]**: Physical Interaction Question Answering tests the understanding of everyday physical processes (Bisk et al., 2020).

- **HellaSwag (HS) [Core]**: Evaluates a model's ability to complete scenarios in a contextually and logically coherent manner (Zellers et al., 2019).

We use the same translations from (Anonymous, 2024). For our evaluations, we use the lm-eval-harness repository[10] for zero-shot accuracy on QA tasks.

B.3.2 Other Evaluation Datasets

- **MMLU**: Multi-domain question answering, MMLU assesses the model's expertise over a wide range of specialized subjects, from professional domains to academia (Hendrycks et al., 2020). We use the human translated versions available from GlobalMMLU (Singh et al., 2024).

- **FrenchBench-MC**: Collection of four evaluations including translated versions of ARC-challenge, HellaSwag, grammar, and vocab (Faysse et al., 2024).

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

Percentile	RPJ2 FR	RPJ2 DE	FW2 FR	FW2 DE	FW2 ZH
95	0.4014				
90	0.2170	0.1654	0.2920	0.2651	0.2751
70	0.0610	0.033931	0.0884	0.0614	0.0722
60	0.0361	0.019172	0.0546	0.0355	0.0437
40	0.0133	0.006802	0.0237	0.0130	0.0170
30	0.0077	0.003946	0.0140	0.0077	0.0107
10	0.0017	0.000884	0.0029	0.0019	0.0028

Table 7: Filter percentile scores for different datasets.

- **Regional**: Evaluation on both the Include (Romanou et al., 2024) and Kaleidoscope (Salazar et al., 2025) benchmarks. For Kaleidoscope, we use only the portion that does not require image modality. As both evaluation sets are small and test regional knowledge, we group both and average the accuracy.

- **NLI**: We report accuracy over French topic-based NLI (Faysse et al., 2024), and XNLI (Conneau et al., 2018) translated into French.

B.4 Licenses and Attributions

The training datasets are supported by public licenses including ODC and Apache license. The pre-trained models including Mistral (for translation), SentenceBert, and OH FastText classifiers are also supported by Apache and MIT licenses. The translated data for Section 3 uses a proprietary translation model following (Seto et al., 2024).

All models and datasets are collected from Huggingface via the datasets library, and all models are evaluated using the lm-eval-harness library from EleutherAI (Gao et al., 2024), which uses an MIT license.

We use the Megatron codebase under the Nvidia license for pre-training.

C Curse of Multilinguality

Early works found that pretraining language models on a large number of languages leads to a decrease in performance for each language (Conneau and Lample, 2019; Rust et al., 2021; Wang et al., 2020; Chang et al., 2024). Several works have investigated causes of performance degradation (Rust et al., 2021; Wang et al., 2020; Chang et al., 2024), and methods for addressing this (Blevins et al., 2024; Pfeiffer et al., 2022). Our work focuses on bilingual language model performance degradation, which limits to a degree the impact of many languages and focuses instead on data size, quality, and training time. While our work can help shed light on factors impacting multilingual model training, our focus is on mitigating performance gaps and the reason for these gaps.

D SBERT Classifier Embeddings for Classification

Prior works have used SBERT for training a linear classifier (Minaee et al., 2021; Albalak et al., 2024; Grangier et al., 2024a), but only in the same language and do not its impact in multilingual LM learning. We show that training a quality classifier with only English data is feasible.

We train a K-means clustering with 64 balanced clusters over the embeddings of 10 files of RedPajamav2 French data to examine the distribution of different datasets following (Grangier et al., 2024b). We then label sets of data in both English and French including C4 (Xue et al., 2021), DCLM classifier training data (Li et al., 2024), and ARC Easy(Clark et al., 2018). Figure 8 shows that both French and English data follow similar histograms indicating that data lie close in the same clusters and can be interchanged when filtering data. As a result we will be able to select the same distribution of data for training models.

E German and English Data Quality Experiments

We start replicating setup in which a bilingual language model is trained on an equal proportion of data from mC4 in German (DE) and English (EN) totaling 100K steps each, following the setup for French and English presented in Section 3.

Table 8 shows performance on MMLU and Core benchmarks. Our findings match those in prior works where we see a 3% drop in English and an

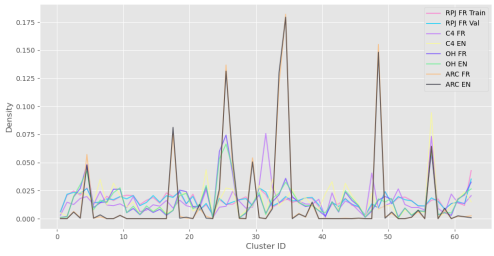


Figure 8: Cluster histograms for distribution of different datasets over RedPajama2. Both English and French versions of the data have similar distributions

Model	Core EN	Core DE	MMLU EN	MMLU DE
EN	**56.2**	40.9	**29.8**	26.4
DE	46.5	48.9	26.8	27.4
BI	53.3	**49.7**	28.9	**28.0**

Table 8: Zero shot accuracy for general understanding and specialized knowledge tasks for monolingual English (EN), German (DE), and bilingual (BI) models.

increase in German of 1% compared to the bilingual model.

Next, we control the data quality and languages in the model. We follow the same setup as in French and vary both the data quality and the language with mC4 and FineWebEDU translations using the same translation system for mC4 and TransWebEDU translations for German.

Figure 9 shows the performance difference when varying quality (y-axis) and language (x-axis) for two sets of zero-shot evaluations: Core and MMLU. When examining the plots, we see that the bottom right square corresponding to a monolingual model trained on high quality in the targeted language has the highest performance. This is closely followed by models which individually vary the language but keep high quality [middle bottom square, e.g., (bi, high)] or mix quality but keep the same language [right middle square, e.g., (EN, mix)]. However, mixed quality and mixed language taken together [middle square, e.g., (bi, mix)] exhibits an average 2.5% drop in English performance by comparison.

Finally, we show results at 30K steps for English and German following the analysis for French. Results are shown in Figure 10. At this scale, we see that both the bilingual high quality (middle bottom) and mixed quality native monolingual (right middle) models have 2.5% lower performance than the monolingual high quality unlike prior results at 200K steps for English evaluations. Similarly low quality results (top row) drop below bilingual again indicating data quality has a large role in training.

F Additional Filter Results

This section presents additional filter percentile results for German and French at larger steps.

F.1 Filter Percentile Results at 200K Steps

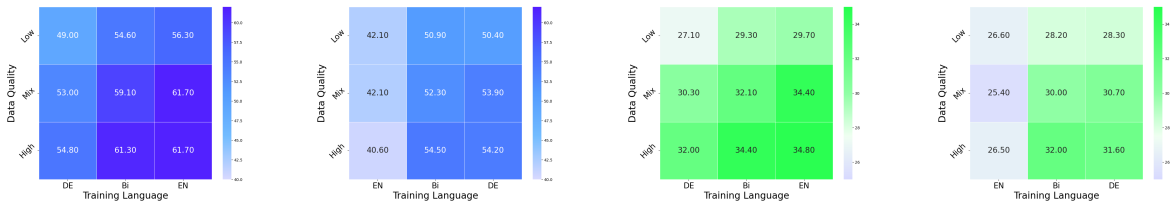
Figure 11 presents results for different filter percentiles at 200K steps for monolingual French models. We see that increasing the percentile used in filtering increases performances on benchmarks. The model performs better than base RedPajamav2 at around 50% quality filter. However, the filtered data models perform worse than training for 200K steps on TransWebEDU on translated benchmarks. Finally, we note that at 95% quality percentile we observe a plateau in performance where the 90th percentile performs better by $\sim 1\%$.

F.2 FineWeb2 French Filter Percentile Results at 200K Steps

Figure 12 presents results for different filter percentiles at 200K steps for monolingual French models. We see that increasing the percentile used in filtering FineWeb2 increases performance on benchmarks. The model performs better than base RedPajamav2 at 30% quality filter, and better than base FineWeb2 at 70% quality filter percentile. The filtered data models achieves the same performance a TransWebEDU on translated benchmarks at the 90th percentile.

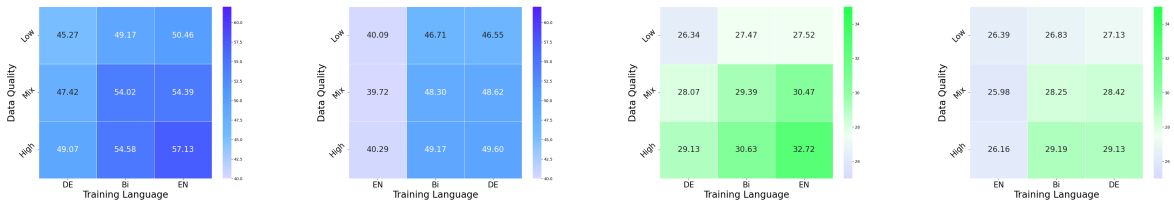
F.3 RedPajamav2 German Filter Percentile Results

Figure 13 presents results for different filter percentiles at 30K steps for monolingual German models. We compare the SBert classifier with training on translated data following the same recipe as for training the original DCLM filter. We see that increasing the percentile used in filtering increases performances on Core tasks across all filters. However, translating with a weak translation system appears to plateau performance, with the small CPU translation system with filtering attaining the same performance as the model with no filtering. Training with a better translations system such as Mistral-7B improves performance to SBert, but requires translating with a more expensive translation system. Second, model trained with SBert filtered data performs better than base RedPajamav2 at around 60% quality filter. However, the filtered data models perform worse than training for 30K



(a) English Core (b) German Core (c) English MMLU (d) German MMLU

Figure 9: Performance with varying data quality and language. Models are trained on combinations of mC4 (low) and FineWebEDU (high) in native English (EN) and translated to German (DE). Models are trained for 200K steps and evaluated on Core (avg over six common-sense reasoning tasks) and MMLU.



(a) English Core (b) German Core (c) English MMLU (d) German MMLU

Figure 10: Performance with varying data quality and language. Models are trained on combinations of mC4 (low) and FineWebEDU (high) in native English (EN) and translated to German (DE). Models are trained for 30K steps and evaluated on Core and MMLU.

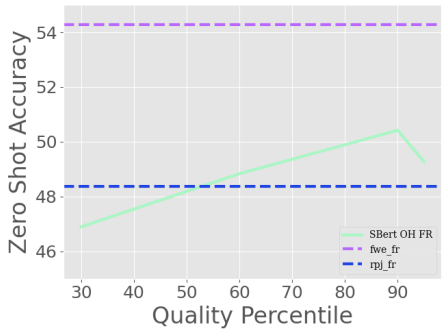


Figure 11: Quality vs. Zero-shot Accuracy on Core tasks for filtered RedPajama2 in French (SBert OH FR) compared with TransWebEDU (fwe_de) and base RedPajamav2 in French (rpj_fr) after 200K steps.

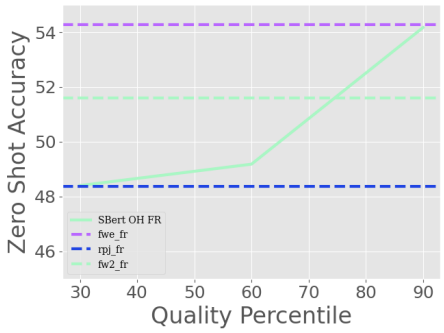


Figure 12: Quality vs. Zero-shot Accuracy on Core tasks for filtered FineWeb2 in French (SBert OH FR) compared with TransWebEDU (fwe_fr), base FineWeb2 in Frnech (fw2_fr), and base RedPajamav2 in French (rpj_fr) after 200K steps.

steps on TransWebEDU (German) on translated benchmarks consistent with our experiments on French.

G Continued Pretraining on High Quality Data

We additionally examine performance when pretraining on the base RedPajamav2 without filtering and subsequently continue pretraining. We experiment with a 1.3B parameter model and examine continuing pretraining after 150K steps, and after 200K steps. We report results in Figure 14. Results indicate that after only a few steps (20-30K) we see performance increase consistent with pretraining on filtered data from scratch indicating computational gains if a pretrained model already exists.

H Results with FWE Training Data for Quality Classifier

Our primary experiments use data from the DCLM classifier training for defining high quality data. However, there may be several definitions of quality. We analyze one possible alternative: textbook quality data as defined by FineWebEDU. We use the same data and annotations used to train the

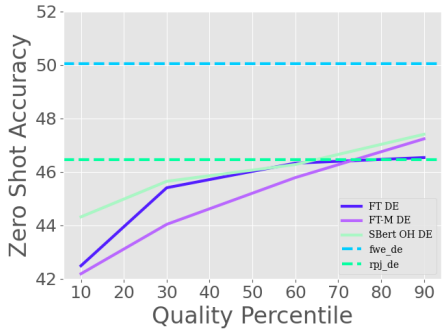


Figure 13: Quality vs. Zero-shot Accuracy on Core tasks for filtered RedPajama2 in German (SBert OH DE) compared with filtering using a FastText classifier trained on translated DCLM classifier training data (FT DE and FT-M DE), TransWebEDU (fwe_de) and base RedPajamav2 in German (rpj_de) after 200K steps.

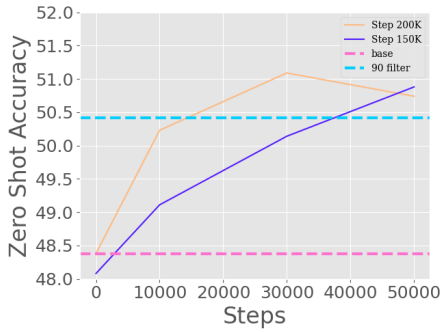


Figure 14: Continued pretraining experiments for 1.3B model continuing training from 150K steps and 200K steps for a total of 50K steps.

FineWebEDU classifier (Penedo et al., 2024a)[11]. We train a binary classifier using the same recipe as prior where the quality label is whether or not the annotation was 2 or above. Our results are in Figure 15.

We find that the DCLM classifier data performs better, while the FineWebEDU data attains the same performance as the base model. There are a few possibilities that we leave to future work: (i) The original FineWebEDU classifier scores between 0 and 5. When training a binary classifier, an accuracy of 82% is achieved for scoring 3 and above as high quality. Decreasing to a 2, might make the task harder, resulting in more low quality examples being selected for training. We chose a classifier score of 2 for high quality as this corresponds to only 30 (ii) Classifying textbook quality data might be a more challenging task when using a universal embedding as the FineWebEDU classifier is a much larger classifier (also an embedding model) than the FastText classifiers. It's possible that with a more complex classifier, and finetuning performance might be higher.

[11]The data is available at https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu-llama3-annotations.

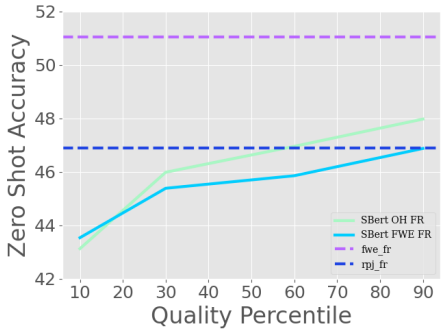


Figure 15: Comparison between training a quality classifier using the DCLM classifier data, and the FineWebEDU data at 30K steps on the Core benchmark datasets for French.

I Comparison with FW2 HQ

We additionally compare with the data selection method of (Messmer et al., 2025). This selection method is similar to the embedding classification approach used in this work, and both build upon the classification method in (Grangier et al., 2024a). However (Grangier et al., 2024a) only study filtering English data and for specific domains. (Messmer et al., 2025) applied the filtering to other languages than English from FineWeb2, and although they aim to filter for high quality data in general, they use datasets such as MMLU and Include, making the filtering aimed more at specialization as in (Grangier et al., 2024a) which also uses MMLU for selection. They are also primarily focused on monolingual performance, and follow the same regime of using data from the target language for training the classifier, and only test monolingual performance. Collectively we refer to their dataset as *FineWeb2 HQ*. We train monolingual and bilingual models for the same 200K steps using the dataset made available[12] for French. We compare with the same percentile of filtering (90%) which should yield approximately the same amount of data. We evaluate on the Core benchmarks in Table 9.

Our findings show similar performance in English, and that our filtering achieves better performance in French, especially for monolingual. Noting that, we are able to achieve better performance

[12]https://huggingface.co/datasets/epfml/FineWeb2-HQ

Model	Core EN	Core FR
FW2	–	51.61
FW2 HQ	–	52.9
(Ours) FW2 90%	–	**54.2**
FW2	60.26	53.65
FW2 HQ	**61.2**	54.9
(Ours) FW2 90%	61.1	**55.5**

Table 9: Performance across Core benchmarks comparing models trained with filtering from our approach and (Messmer et al., 2025) on FineWeb2. Top rows represent monolingual performance, and bottom are multilingual with FineWebEDU as the English data.

without access to high quality data from other languages for training[13]. We hypothesize that this may be because using data from MMLU and Include are very task specific, rather than general high quality data. Include has specific regional knowledge such as driving exams, which may be filtered for instead, and MMLU is predominantly science knowledge and Western (possibly American) culturally influenced. Thus, there may be similar issues as we found with the FineWebEDU training annotation set in Section H.

J German and Chinese Filtering Evaluations

Section 5 focuses on French-English bilingual models due to the (i) availability of a sufficient amount of data for training 1.3B models for both curated data (FineWeb2) and common crawl data (Redpajama2), (ii) multiple evaluations both translated from English data and native, and (iii) closeness to English. In this section, we additionally show that the filtering improves performance for other languages in both monolingual and bilingual models.

We compare results for French, German, and Chinese languages from FineWeb2 in Table 10 for monolingual models and Table 11 for bilingual models, and French and German for RedPajamav2[14]. In all cases, we observe an improvement in performance with French and German models being better than training on translated high quality English data. Note further that for all three models, there is reduced gap in performance between monolingual and bilingual models for all languages. We report that the gap in English performances to a monolingual English model improves by up to ∼1% with filtered data from FineWeb2 in other languages.

Model	RPJ2 FR	RPJ2 DE	FW2 FR	FW2 DE	FW2 ZH
Base	48.38	48.33	51.61	49.83	51.36
90% Filter	50.42	49.86	54.17	52.53	53.14

Table 10: Comparison of filtering for RPJ2 and FW2 for monolingual models for Core benchmarks in the native languages.

Model	RPJ2 FR	RPJ2 DE	FW2 FR	FW2 DE	FW2 ZH
Base	52.16	50.70	53.65	52.25	51.50
90% Filter	53.37	51.58	55.47	53.90	53.97

Table 11: Comparison of filtering for RPJ2 and FW2 for bilingual models for Core benchmarks in the non-english languages. Models are trained with the respective datasets and FineWebEDU in English.

K Model Scaling

K.1 Model Scaling Experiments

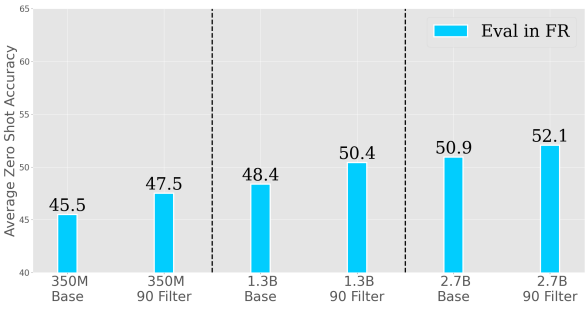


Figure 16: Monolingual model performance comparing filtering on the Core FR benchmarks for various model sizes.

Methodology: We investigate to what extent our results are similar across model sizes. We measure performance at three model sizes: 350M, 1.3B, and 2.7B, and train models for each model size trained on the same data pools for both the base distribution of RedPajamav2 FR and filtered version at 90% filtering. Note that the 2.7B model has twice the context length and sees data for twice as many repetitions. We report results for the monolingual models in Figure 16 and for bilingual performance in Figure 17.

[13]Note that we did not conduct full comparison to other tasks in our list of benchmarks as both MMLU and Regional evaluations have data used for selection

[14]For German experiments on RedPajamav2, we use a smaller set of data and repeat for two repetitions. This amount of repetition should not have an effect following (Muennighoff et al., 2024). Models get similar performance and improvements as for RedPajamav2 in French.

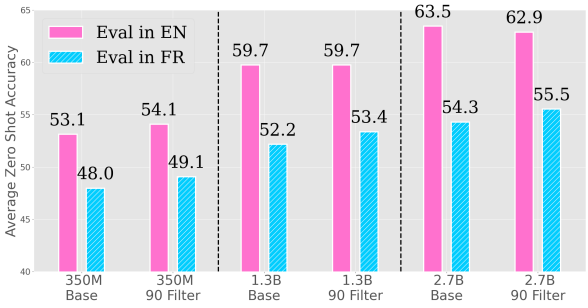


Figure 17: Bilingual model performance comparing filtering on the Core EN and FR benchmarks for various model sizes.

Findings: For monolingual models, we see 2% improvement for 350M and 1.3B, and 1% for the 2.7B model. With more data, it's possible to see greater improvements on the 2.7B model, however we note that we both saturate on the amount of filtered data requiring multiple repetitions, and the scale drops to only 2-3× Chinchilla vs. 7×. For bilingual performance, we see consistent performance of 1-1.5% improvement for French evaluations. However, there is little improvement from filtering on English performance as is similar to prior results. This is similar to results in Section 5.5, and is likely a result of the English FineWebEDU dataset having higher quality data relevant to the downstream evaluations than even filtered RedPajamav2.

K.2 Comparison for 2.7B Parameter Models

Section K.1 shows that as we increase model size, performance also increases indicating that filtering improves performance regardless of model size. We first show that this also applies at all intermediate checkpoints where we observe consistent trends regardless of the number of steps in Figure 18. Next, we show comparison with public state-of-the-art multilingual models such as Qwen2.5 3B (Yang et al., 2024), and Helium-1[15] in Table 12 on the same French evaluation sets from Section 5.6. At the 2.7B parameter scale, our model trained with filtered RedPajamav2 outperforms the unfiltered model. We do not train on the FineWeb2 as the amount of filtered data is small and would require over 10x repetition which may impact performance. Performance of our models are lower than Qwen2.5 and Helium-1 models at this scale. This is because our models are trained with the same data as the smaller models for consistency and comparison leading to more repetitions even for English data, and models such as the Qwen2.5 family are trained on 18T tokens (Yang et al., 2024), which is over 40× the data used by our models. We see consistent improvements in early stages of training (first repetition of data), and expect that with more data in other languages for filtering, performance could improve as well.

[15] https://kyutai.org/2025/04/30/helium.html

L Individual Task Accuracies

We provide individual accuracies for all models we train with our filtering strategy and evaluate in Section 5.

L.1 Monolingual French Filter Performance

This section provides results for individual tasks in the Core benchmark for monolingual French models with varying quality to supplement Figure 5, and Table 4. Results are presented in Table 13.

L.2 Bilingual French Filter Performance

This section provides results for individual tasks in the Core benchmark for monolingual French models with varying quality to supplement Figure 6. Results are presented in Table 14.

L.3 Filter Performance Comparison with FineWeb 2 HQ

We report performance for individual tasks comparing FineWeb2 with our filtering and FineWeb2 HQ (Messmer et al., 2025). Table 15 shows results for monolingual models and Table 16 for bilingual models with FineWebEDU in English.

L.4 Filter Performance Across Languages

This section expands on results for Tables 10-11. For monolingual models, Table 17 presents individual task accuracy for French, Table 18 for German, and Table 19 for Chinese.

individual task accuracies for bilingual models with FineWebEDU in English are included in Table 20 for French, Table 21 for German, and Table 22 for Chinese.

L.5 Filter Performance for Varying Model Sizes

We expand our results for Figure 17 showing accuracy for individual tasks in the Core tasks for different model sizes for bilingual English-French models. All models are trained with FineWebEDU (FWE) as the English data and RedPajama2 (RPJ2)

Model	Core EN	MMLU EN	Core	MMLU	FB-MC	Regional	NLI	AVG
RPJ2	63.45	35.86	54.30	30.28	61.33	34.54	40.86	44.26
Helium-1 2B	71.26	41.11	60.13	33.53	**64.28**	34.51	42.15	46.92
Bloom 3B	58.21	31.68	52.62	28.93	60.24	31.55	40.50	42.77
Qwen 3B	**72.54**	**44.79**	**60.95**	34.96	63.52	**36.62**	**52.32**	**49.67**
RPJ2 90%	62.89	35.03	55.52	30.72	63.26	34.51	40.70	44.94

Table 12: Comparison of different public 2B+ bilingual models in French and English.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WG	AVG
RPJ2 30%	22.18 ± 1.21	35.06 ± 0.98	30.81 ± 0.46	56.37 ± 1.16	57.30 ± 1.56	51.54 ± 1.40	42.21
RPJ2 60%	23.98 ± 1.25	37.25 ± 0.99	29.71 ± 0.46	57.02 ± 1.16	66.10 ± 1.50	50.36 ± 1.41	44.07
RPJ2 90%	24.40 ± 1.26	37.67 ± 0.99	31.69 ± 0.46	56.42 ± 1.16	61.90 ± 1.54	50.75 ± 1.41	43.80
RPJ2 95%	23.46 ± 1.24	37.92 ± 1.00	30.88 ± 0.46	57.89 ± 1.15	59.10 ± 1.56	50.51 ± 1.41	43.29
FW2 30%	22.95 ± 1.23	37.42 ± 0.99	33.89 ± 0.47	62.30 ± 1.13	62.40 ± 1.53	51.70 ± 1.40	45.11
FW2 60%	24.66 ± 1.26	38.05 ± 1.00	34.90 ± 0.48	61.97 ± 1.13	60.70 ± 1.55	50.04 ± 1.41	45.05
FW2 90%	26.79 ± 1.29	43.43 ± 1.02	38.89 ± 0.49	63.33 ± 1.12	65.80 ± 1.50	52.57 ± 1.40	48.47
RPJ2 30%	25.72 ± 1.29	37.34 ± 1.02	42.16 ± 0.51	63.33 ± 1.12	60.86 ± 1.58	51.93 ± 1.43	46.89
RPJ2 60%	27.03 ± 1.31	42.36 ± 1.04	43.02 ± 0.51	63.38 ± 1.12	64.74 ± 1.55	52.43 ± 1.43	48.83
RPJ2 90%	28.51 ± 1.33	45.84 ± 1.05	45.05 ± 0.51	65.13 ± 1.11	65.06 ± 1.55	52.92 ± 1.43	50.42
RPJ2 95%	27.29 ± 1.32	44.52 ± 1.04	44.27 ± 0.51	64.85 ± 1.11	61.18 ± 1.58	53.50 ± 1.43	49.27
FW2 30%	25.02 ± 1.28	39.41 ± 1.03	45.20 ± 0.52	66.05 ± 1.10	60.65 ± 1.58	54.07 ± 1.43	48.40
FW2 60%	27.55 ± 1.32	39.63 ± 1.03	47.41 ± 0.52	66.32 ± 1.10	61.18 ± 1.58	53.00 ± 1.43	49.18
FW2 90%	30.95 ± 1.37	48.88 ± 1.05	52.55 ± 0.52	68.77 ± 1.08	68.31 ± 1.51	55.56 ± 1.43	54.17

Table 13: Evaluation of 1.3B parameter monolingual French models on general understanding tasks for English (top) and French (bottom) with varying quality. All evaluations are zero-shot.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WG	AVG
FWE FR	32.00 ± 1.36	53.45 ± 1.02	42.53 ± 0.49	65.23 ± 1.11	72.10 ± 1.42	54.78 ± 1.40	53.35
FWE EN	37.71 ± 1.42	66.08 ± 0.97	56.84 ± 0.49	73.78 ± 1.03	81.20 ± 1.24	57.38 ± 1.39	62.16
FWE EN RPJ2 30%	32.76 ± 1.37	62.08 ± 1.00	51.66 ± 0.50	71.33 ± 1.06	78.80 ± 1.29	57.14 ± 1.39	58.96
FWE EN RPJ2 60%	33.28 ± 1.38	62.29 ± 0.99	52.09 ± 0.50	72.09 ± 1.05	81.50 ± 1.23	56.43 ± 1.39	59.61
FWE EN RPJ2 90%	34.90 ± 1.39	62.46 ± 0.99	52.75 ± 0.50	71.33 ± 1.06	80.70 ± 1.25	56.27 ± 1.39	59.73
FWE EN FW2 30%	34.13 ± 1.39	60.40 ± 1.00	53.07 ± 0.50	72.42 ± 1.04	79.30 ± 1.28	56.99 ± 1.39	59.38
FWE EN FW2 60%	33.53 ± 1.38	61.53 ± 1.00	54.48 ± 0.50	72.31 ± 1.04	80.10 ± 1.26	56.12 ± 1.39	59.68
FWE EN FW2 90%	36.26 ± 1.40	62.37 ± 0.99	55.74 ± 0.50	73.50 ± 1.03	81.30 ± 1.23	57.22 ± 1.39	61.07
FWE FR	34.26 ± 1.40	53.28 ± 1.05	48.13 ± 0.52	63.17 ± 1.13	70.41 ± 1.48	56.46 ± 1.42	54.28
FWE EN	25.11 ± 1.28	32.94 ± 0.99	30.64 ± 0.48	52.01 ± 1.17	62.22 ± 1.57	50.95 ± 1.43	42.31
FWE EN RPJ2 30%	29.56 ± 1.35	45.88 ± 1.05	46.13 ± 0.52	63.66 ± 1.12	66.84 ± 1.53	53.58 ± 1.43	50.94
FWE EN RPJ2 60%	30.95 ± 1.37	47.95 ± 1.05	47.44 ± 0.52	66.00 ± 1.11	67.58 ± 1.52	54.73 ± 1.43	52.44
FWE EN RPJ2 90%	31.12 ± 1.37	49.45 ± 1.05	48.89 ± 0.52	65.94 ± 1.11	69.57 ± 1.49	55.23 ± 1.43	53.37
FWE EN FW2 30%	29.99 ± 1.35	45.40 ± 1.04	47.86 ± 0.52	66.54 ± 1.10	67.47 ± 1.52	53.91 ± 1.43	51.86
FWE EN FW2 60%	28.86 ± 1.34	45.31 ± 1.04	50.07 ± 0.52	66.70 ± 1.10	67.47 ± 1.52	55.80 ± 1.43	52.37
FWE EN FW2 90%	32.61 ± 1.38	50.15 ± 1.05	53.21 ± 0.52	69.26 ± 1.08	71.14 ± 1.47	56.46 ± 1.42	55.47

Table 14: Evaluation of 1.3B parameter models on general understanding tasks for English (top) and French (bottom) with varying quality. All evaluations are zero-shot.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WG	AVG
FW2 FR	25.77 ± 1.28	39.44 ± 1.00	37.21 ± 0.48	64.09 ± 1.12	65.90 ± 1.50	52.41 ± 1.40	47.47
FW2 FR 90%	26.79 ± 1.29	43.43 ± 1.02	38.89 ± 0.49	63.33 ± 1.12	65.80 ± 1.50	52.57 ± 1.40	48.47
FW2 FR HQ	27.82 ± 1.31	44.19 ± 1.02	37.40 ± 0.48	63.55 ± 1.12	69.10 ± 1.46	52.25 ± 1.40	49.05
FW2 FR	28.95 ± 1.34	43.28 ± 1.04	48.58 ± 0.52	67.85 ± 1.09	66.84 ± 1.53	54.16 ± 1.43	51.61
FW2 FR 90FW2 FR HQ	30.69 ± 1.36	46.46 ± 1.05	49.87 ± 0.52	68.17 ± 1.09	67.58 ± 1.52	54.90 ± 1.43	52.94

Table 15: Evaluation of 1.3B parameter models on general understanding tasks for English (top) and French (bottom) with varying quality. Models are trained on FineWeb2 (FW2) or FineWeb 2 HQ (FW2 HQ) in French with and without filtering. All evaluations are zero-shot.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WG	AVG
FWE EN FW2 FR	34.64 ± 1.39	62.25 ± 0.99	54.31 ± 0.50	72.31 ± 1.04	80.30 ± 1.26	57.77 ± 1.39	60.26
FWE EN + FW2 FR 90%	36.26 ± 1.40	62.37 ± 0.99	55.74 ± 0.50	73.50 ± 1.03	81.30 ± 1.23	57.22 ± 1.39	61.07
FWE EN FW2 FR HQ	35.92 ± 1.40	64.35 ± 0.98	54.57 ± 0.50	73.34 ± 1.03	81.80 ± 1.22	57.22 ± 1.39	61.20
FWE EN FW2 FR	31.21 ± 1.37	48.75 ± 1.05	50.27 ± 0.52	68.12 ± 1.09	70.30 ± 1.48	53.25 ± 1.43	53.65
FWE EN + FW2 FR 90%	32.61 ± 1.38	50.15 ± 1.05	53.21 ± 0.52	69.26 ± 1.08	71.14 ± 1.47	56.46 ± 1.42	55.47
FWE EN FW2 FR HQ	32.52 ± 1.38	51.74 ± 1.05	51.31 ± 0.52	68.55 ± 1.08	71.46 ± 1.46	53.50 ± 1.43	54.85

Table 16: Evaluation of 1.3B parameter models on general understanding tasks for English (top) and French (bottom) with varying quality. Models are trained on FineWeb2 (FW2) or FineWeb 2 HQ (FW2 HQ) in French with and without filtering and FineWebEDU (FWE) in English. All evaluations are zero-shot.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WG	AVG
RPJ2 FR	23.46 ± 1.24	36.78 ± 0.99	32.03 ± 0.47	56.86 ± 1.16	64.80 ± 1.51	49.96 ± 1.41	43.98
RPJ2 FR 90%	24.40 ± 1.26	37.67 ± 0.99	31.69 ± 0.46	56.42 ± 1.16	61.90 ± 1.54	50.75 ± 1.41	43.80
FW2 FR	25.77 ± 1.28	39.44 ± 1.00	37.21 ± 0.48	64.09 ± 1.12	65.90 ± 1.50	52.41 ± 1.40	47.47
FW2 FR 90%	26.79 ± 1.29	43.43 ± 1.02	38.89 ± 0.49	63.33 ± 1.12	65.80 ± 1.50	52.57 ± 1.40	48.47
RPJ2 FR	24.85 ± 1.28	41.35 ± 1.03	43.52 ± 0.51	64.04 ± 1.12	62.54 ± 1.57	53.99 ± 1.43	48.38
RPJ2 FR 90%	28.51 ± 1.33	45.84 ± 1.05	45.05 ± 0.51	65.13 ± 1.11	65.06 ± 1.55	52.92 ± 1.43	50.42
FW2 FR	28.95 ± 1.34	43.28 ± 1.04	48.58 ± 0.52	67.85 ± 1.09	66.84 ± 1.53	54.16 ± 1.43	51.61
FW2 FR 90%	30.95 ± 1.37	48.88 ± 1.05	52.55 ± 0.52	68.77 ± 1.08	68.31 ± 1.51	55.56 ± 1.43	54.17

Table 17: Evaluation of 1.3B parameter models on general understanding tasks for English (top) and French (bottom) with varying quality. Models are trained on RedPajamav2 (RPJ2) or FineWeb2 (FW2) in French with and without filtering. All evaluations are zero-shot.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WG	AVG
RPJ2 DE	24.23 ± 1.25	37.92 ± 1.00	31.69 ± 0.46	58.16 ± 1.15	63.30 ± 1.52	49.64 ± 1.41	44.16
RPJ2 DE 90%	24.49 ± 1.26	38.59 ± 1.00	31.53 ± 0.46	58.22 ± 1.15	64.70 ± 1.51	51.38 ± 1.40	44.82
FW2 DE	24.91 ± 1.26	35.10 ± 0.98	36.58 ± 0.48	61.70 ± 1.13	67.10 ± 1.49	52.01 ± 1.40	46.23
FW2 DE 90%	24.57 ± 1.26	42.59 ± 1.01	37.86 ± 0.48	63.44 ± 1.12	66.70 ± 1.49	50.28 ± 1.41	47.57
RPJ2 DE	27.88 ± 1.33	42.74 ± 1.04	40.02 ± 0.51	61.64 ± 1.13	67.89 ± 1.52	52.28 ± 1.45	48.74
RPJ2 DE 90%	29.99 ± 1.36	45.22 ± 1.05	41.29 ± 0.51	63.06 ± 1.13	71.16 ± 1.47	51.27 ± 1.45	50.33
FW2 DE	27.79 ± 1.33	39.07 ± 1.03	43.58 ± 0.51	64.20 ± 1.12	69.79 ± 1.49	54.56 ± 1.45	49.83
FW2 DE 90%	29.11 ± 1.35	47.48 ± 1.05	47.78 ± 0.52	66.16 ± 1.10	69.26 ± 1.50	55.41 ± 1.45	52.53

Table 18: Evaluation of 1.3B parameter models on general understanding tasks for English (top) and German (bottom) with varying quality. Models are trained on RedPajamav2 (RPJ2) or FineWeb2 (FW2) in German with and without filtering. All evaluations are zero-shot.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WG	AVG
FW2 ZH	22.18 ± 1.21	36.62 ± 0.99	30.09 ± 0.46	57.07 ± 1.15	64.20 ± 1.52	52.41 ± 1.40	43.76
FW2 ZH 90%	21.76 ± 1.21	36.36 ± 0.99	30.73 ± 0.46	58.16 ± 1.15	64.30 ± 1.52	53.04 ± 1.40	44.06
FW2 ZH	28.62 ± 1.34	47.56 ± 1.05	40.25 ± 0.51	61.92 ± 1.13	76.00 ± 1.35	53.82 ± 1.53	51.36
FW2 ZH 90%	30.02 ± 1.35	52.27 ± 1.05	43.05 ± 0.51	63.33 ± 1.12	77.60 ± 1.32	52.60 ± 1.54	53.14

Table 19: Evaluation of 1.3B parameter models on general understanding tasks for English (top) and Chinese (bottom) with varying quality. Models are trained on RedPajamav2 (RPJ2) or FineWeb2 (FW2) in Chinese with and without filtering. All evaluations are zero-shot.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WG	AVG
FWE EN RPJ2 FR	35.32 ± 1.40	61.57 ± 1.00	52.56 ± 0.50	70.73 ± 1.06	81.60 ± 1.23	56.67 ± 1.39	59.74
FWE EN RPJ2 FR 90%	34.90 ± 1.39	62.46 ± 0.99	52.75 ± 0.50	71.33 ± 1.06	80.70 ± 1.25	56.27 ± 1.39	59.73
FWE EN FW2 FR	34.64 ± 1.39	62.25 ± 0.99	54.31 ± 0.50	72.31 ± 1.04	80.30 ± 1.26	57.77 ± 1.39	60.26
FWE EN + FW2 FR 90%	36.26 ± 1.40	62.37 ± 0.99	55.74 ± 0.50	73.50 ± 1.03	81.30 ± 1.23	57.22 ± 1.39	61.07
FWE EN RPJ2 FR	30.34 ± 1.36	47.16 ± 1.05	47.37 ± 0.52	64.58 ± 1.12	69.67 ± 1.49	53.83 ± 1.43	52.16
FWE EN RPJ2 FR 90%	31.12 ± 1.37	49.45 ± 1.05	48.89 ± 0.52	65.94 ± 1.11	69.57 ± 1.49	55.23 ± 1.43	53.37
FWE EN FW2 FR	31.21 ± 1.37	48.75 ± 1.05	50.27 ± 0.52	68.12 ± 1.09	70.30 ± 1.48	53.25 ± 1.43	53.65
FWE EN + FW2 FR 90%	32.61 ± 1.38	50.15 ± 1.05	53.21 ± 0.52	69.26 ± 1.08	71.14 ± 1.47	56.46 ± 1.42	55.47

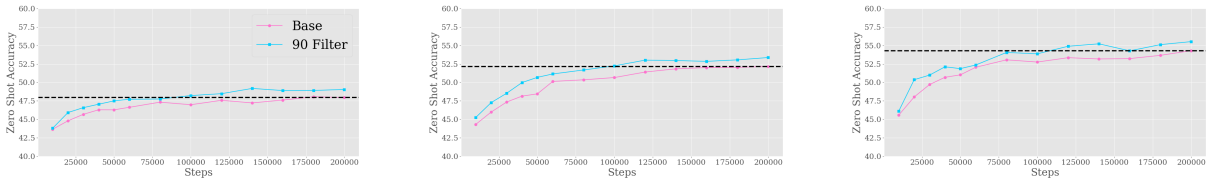
Table 20: Evaluation of 1.3B parameter models on general understanding tasks for English (top) and French (bottom) with varying quality. Models are trained on FineWebEDU (FWE) in English and RedPajamav2 (RPJ2) or FineWeb2 (FW2) in French with and without filtering. All evaluations are zero-shot.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WG	AVG
FWE EN RPJ2 DE	35.32 ± 1.40	61.53 ± 1.00	51.43 ± 0.50	70.62 ± 1.06	82.00 ± 1.22	54.85 ± 1.40	59.29
FWE EN RPJ2 DE 90%	33.70 ± 1.38	61.95 ± 1.00	51.36 ± 0.50	70.78 ± 1.06	82.30 ± 1.21	55.49 ± 1.40	59.26
FWE EN FW2 DE	33.28 ± 1.38	59.60 ± 1.01	53.17 ± 0.50	72.20 ± 1.05	80.70 ± 1.25	56.75 ± 1.39	59.28
FWE EN FW2 DE 90%	35.84 ± 1.40	63.89 ± 0.99	54.93 ± 0.50	71.55 ± 1.05	82.80 ± 1.19	56.12 ± 1.39	60.85
FWE EN RPJ2 DE	29.29 ± 1.35	46.90 ± 1.05	42.49 ± 0.51	62.13 ± 1.13	71.16 ± 1.47	53.38 ± 1.45	50.89
FWE EN RPJ2 DE 90%	30.17 ± 1.36	49.69 ± 1.05	43.95 ± 0.51	62.89 ± 1.13	72.00 ± 1.46	53.21 ± 1.45	51.98
FWE EN FW2 DE	29.46 ± 1.35	47.35 ± 1.05	45.20 ± 0.51	65.45 ± 1.11	72.00 ± 1.46	54.05 ± 1.45	52.25
FWE EN FW2 DE 90%	30.61 ± 1.37	52.30 ± 1.05	48.54 ± 0.52	65.18 ± 1.11	73.47 ± 1.43	53.29 ± 1.45	53.90

Table 21: Evaluation of 1.3B parameter models on general understanding tasks for English (top) and German (bottom) with varying quality. Models are trained on FineWebEDU (FWE) in English and RedPajamav2 (RPJ2) or FineWeb2 (FW2) in German with and without filtering. All evaluations are zero-shot.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WG	AVG
FWE EN FW2 ZH	33.96 ± 1.38	60.98 ± 1.00	49.10 ± 0.50	70.78 ± 1.06	77.80 ± 1.31	55.72 ± 1.40	58.06
FWE EN FW2 ZH 90%	32.76 ± 1.37	62.04 ± 1.00	50.32 ± 0.50	70.62 ± 1.06	80.70 ± 1.25	55.56 ± 1.40	58.67
FWE EN FW2 ZH	29.23 ± 1.34	49.14 ± 1.05	40.15 ± 0.51	61.75 ± 1.13	78.30 ± 1.30	50.42 ± 1.54	51.50
FWE EN FW2 ZH 90%	29.93 ± 1.35	54.34 ± 1.05	43.46 ± 0.51	63.55 ± 1.12	79.40 ± 1.28	53.16 ± 1.53	53.97

Table 22: Evaluation of 1.3B parameter models on general understanding tasks for English (top) and Chinese (bottom) with varying quality. Models are trained on FineWebEDU (FWE) in English and FineWeb2 (FW2) in Chinese with and without filtering. All evaluations are zero-shot.



(a) 350M (b) 1.3B (c) 2.7B

Figure 18: Performance at intermediate checkpoints during training for different model sizes models for Core FR benchmarks.

as the French data with and without filtering. Results are shown in Table 23.

L.6 Comparisons for Public Models

In this section, we provide individual accuracies for all tasks and models in Table 3 and 12. Table 24 shows results for Core tasks. Table 25 shows results for the FrenchBench multiple choice tasks. Table 26 shows results for both the regional knowledge tasks, and the NLI tasks.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WG	AVG
350M FWE EN RPJ2 FR	27.47 ± 1.30	53.83 ± 1.02	43.02 ± 0.49	67.08 ± 1.10	75.00 ± 1.37	52.33 ± 1.40	53.12
350M FWE EN RPJ2 FR 90%	29.18 ± 1.33	56.90 ± 1.02	42.92 ± 0.49	68.12 ± 1.09	75.40 ± 1.36	52.01 ± 1.40	54.09
1.3B FWE EN RPJ2 FR	35.32 ± 1.40	61.57 ± 1.00	52.56 ± 0.50	70.73 ± 1.06	81.60 ± 1.23	56.67 ± 1.39	59.74
1.3B FWE EN RPJ2 FR 90%	34.90 ± 1.39	62.46 ± 0.99	52.75 ± 0.50	71.33 ± 1.06	80.70 ± 1.25	56.27 ± 1.39	59.73
2.7B FWE EN RPJ2 FR	38.65 ± 1.42	68.60 ± 0.95	57.82 ± 0.49	73.12 ± 1.03	84.20 ± 1.15	58.33 ± 1.39	63.45
2.7B FWE EN RPJ2 FR 90%	38.23 ± 1.42	66.08 ± 0.97	57.04 ± 0.49	73.18 ± 1.03	83.60 ± 1.17	59.19 ± 1.38	62.89
350M FWE EN RPJ2 FR	26.85 ± 1.31	42.23 ± 1.04	39.69 ± 0.51	62.02 ± 1.13	64.43 ± 1.55	52.59 ± 1.43	47.97
350M FWE EN RPJ2 FR 90%	28.25 ± 1.33	45.22 ± 1.04	40.93 ± 0.51	62.95 ± 1.13	64.22 ± 1.55	52.76 ± 1.43	49.05
1.3B FWE EN RPJ2 FR	30.34 ± 1.36	47.16 ± 1.05	47.37 ± 0.52	64.58 ± 1.12	69.67 ± 1.49	53.83 ± 1.43	52.16
1.3B FWE EN RPJ2 FR 90%	31.12 ± 1.37	49.45 ± 1.05	48.89 ± 0.52	65.94 ± 1.11	69.57 ± 1.49	55.23 ± 1.43	53.37
2.7B FWE EN RPJ2 FR	32.61 ± 1.38	49.76 ± 1.05	51.34 ± 0.52	67.25 ± 1.09	70.09 ± 1.48	54.73 ± 1.43	54.30
2.7B FWE EN RPJ2 FR 90%	33.65 ± 1.40	51.30 ± 1.05	52.61 ± 0.52	67.90 ± 1.09	71.77 ± 1.46	55.88 ± 1.43	55.52

Table 23: Evaluation of models on general understanding tasks for English (top) and French (bottom) with varying quality. Models are trained on FineWebEDU (FWE) in English and Redpajama2 (RPJ2) in French with and without filtering at varying model sizes. All evaluations are zero-shot.

Model Name	ARC-C	ARC-E	HS	PIQA	SCIQ	WG	AVG
1.3B FWE EN RPJ2 FR	35.32 ± 1.40	61.57 ± 1.00	52.56 ± 0.50	70.73 ± 1.06	81.60 ± 1.23	56.67 ± 1.39	59.74
1.3B FWE EN FW2 FR	34.64 ± 1.39	62.25 ± 0.99	54.31 ± 0.50	72.31 ± 1.04	80.30 ± 1.26	57.77 ± 1.39	60.26
1.3B FWE EN FWE FR	37.29 ± 1.41	64.06 ± 0.98	55.22 ± 0.50	73.45 ± 1.03	83.40 ± 1.18	57.70 ± 1.39	61.85
1.3B FWE EN RPJ 2 90%	34.90 ± 1.39	62.46 ± 0.99	52.75 ± 0.50	71.33 ± 1.06	80.70 ± 1.25	56.27 ± 1.39	59.73
1.3B FWE EN FW2 FR 90%	36.26 ± 1.40	62.37 ± 0.99	55.74 ± 0.50	73.50 ± 1.03	81.30 ± 1.23	57.22 ± 1.39	61.07
2.7B FWE EN RPJ2 FR	38.65 ± 1.42	68.60 ± 0.95	57.82 ± 0.49	73.12 ± 1.03	84.20 ± 1.15	58.33 ± 1.39	63.45
2.7B FWE EN RPJ2 FR 90%	38.23 ± 1.42	66.08 ± 0.97	57.04 ± 0.49	73.18 ± 1.03	83.60 ± 1.17	59.19 ± 1.38	62.89
CroissantLLM	27.56 ± 1.31	52.27 ± 1.02	53.53 ± 0.50	71.60 ± 1.05	79.40 ± 1.28	55.64 ± 1.40	56.67
TransWebLLM	36.18 ± 1.40	62.21 ± 0.99	52.32 ± 0.50	70.51 ± 1.06	80.20 ± 1.26	56.27 ± 1.39	59.61
Bloom 1B	25.68 ± 1.28	45.45 ± 1.02	42.98 ± 0.49	67.14 ± 1.10	74.20 ± 1.38	54.93 ± 1.40	51.73
Qwen2.5 1.5B	45.14 ± 1.45	71.46 ± 0.93	67.75 ± 0.47	76.06 ± 1.00	92.90 ± 0.81	63.38 ± 1.35	69.45
EuroLLM 1.7B	37.46 ± 1.41	64.10 ± 0.98	59.38 ± 0.49	73.45 ± 1.03	84.90 ± 1.13	59.04 ± 1.38	63.05
Helium-1 2B	46.50 ± 1.46	73.74 ± 0.90	69.63 ± 0.46	78.62 ± 0.96	92.30 ± 0.84	66.77 ± 1.32	71.26
Bloom 3B	30.55 ± 1.35	53.24 ± 1.02	54.51 ± 0.50	70.51 ± 1.06	81.70 ± 1.22	58.72 ± 1.38	58.21
Qwen2.5 3B	47.44 ± 1.46	73.11 ± 0.91	73.53 ± 0.44	78.84 ± 0.95	93.80 ± 0.76	68.51 ± 1.31	72.54
1.3B FWE EN RPJ2 FR	30.34 ± 1.36	47.16 ± 1.05	47.37 ± 0.52	64.58 ± 1.12	69.67 ± 1.49	53.83 ± 1.43	52.16
1.3B FWE EN FW2 FR	31.21 ± 1.37	48.75 ± 1.05	50.27 ± 0.52	68.12 ± 1.09	70.30 ± 1.48	53.25 ± 1.43	53.65
1.3B FWE EN FWE FR	33.39 ± 1.39	53.46 ± 1.05	47.68 ± 0.52	62.57 ± 1.13	72.40 ± 1.45	55.39 ± 1.43	54.15
1.3B FWE EN RPJ 2 90%	31.12 ± 1.37	49.45 ± 1.05	48.89 ± 0.52	65.94 ± 1.11	69.57 ± 1.49	55.23 ± 1.43	53.37
1.3B FWE EN FW2 FR 90%	32.61 ± 1.38	50.15 ± 1.05	53.21 ± 0.52	69.26 ± 1.08	71.14 ± 1.47	56.46 ± 1.42	55.47
2.7B FWE EN RPJ2 FR	32.61 ± 1.38	49.76 ± 1.05	51.34 ± 0.52	67.25 ± 1.09	70.09 ± 1.48	54.73 ± 1.43	54.30
2.7B FWE EN RPJ2 FR 90%	33.65 ± 1.40	51.30 ± 1.05	52.61 ± 0.52	67.90 ± 1.09	71.77 ± 1.46	55.88 ± 1.43	55.52
CroissantLLM	27.90 ± 1.32	45.22 ± 1.04	50.52 ± 0.52	66.87 ± 1.10	69.67 ± 1.49	55.31 ± 1.43	52.58
TransWebLLM	34.79 ± 1.41	53.68 ± 1.05	48.21 ± 0.52	63.76 ± 1.12	75.45 ± 1.39	54.40 ± 1.43	55.05
Bloom 1B	27.03 ± 1.31	40.03 ± 1.03	41.56 ± 0.51	61.70 ± 1.13	67.16 ± 1.52	54.73 ± 1.43	48.70
Qwen2.5 1.5B	32.69 ± 1.39	51.12 ± 1.05	49.63 ± 0.52	63.06 ± 1.13	79.54 ± 1.31	57.86 ± 1.42	55.65
EuroLLM 1.7B	31.39 ± 1.37	51.17 ± 1.05	51.47 ± 0.52	65.18 ± 1.11	74.08 ± 1.42	56.38 ± 1.42	54.94
Helium-1 2B	36.09 ± 1.42	55.00 ± 1.04	59.51 ± 0.51	67.90 ± 1.09	81.64 ± 1.25	60.66 ± 1.40	60.13
Bloom 3B	30.08 ± 1.35	45.49 ± 1.05	51.04 ± 0.52	65.13 ± 1.11	69.78 ± 1.49	54.24 ± 1.43	52.62
Qwen2.5 3B	38.19 ± 1.44	55.70 ± 1.04	58.58 ± 0.51	65.18 ± 1.11	84.58 ± 1.17	63.46 ± 1.38	60.95

Table 24: Evaluation of our models against public models on Core "general understanding" tasks for English (top) and French (bottom). All evaluations are zero-shot.

Model Name	ARC-C	Grammar	HS	Vocab
1.3B FWE EN RPJ2 FR	30.54 ± 1.35	82.35 ± 3.51	47.4 ± 0.52	80.67 ± 3.64
1.3B FWE EN FW2 FR	31.05 ± 1.35	80.67 ± 3.64	50.26 ± 0.52	78.99 ± 3.75
1.3B FWE EN FWE FR	36.7 ± 1.41	68.07 ± 4.29	47.59 ± 0.52	65.55 ± 4.37
1.3B FWE EN RPJ 2 90%	32.34 ± 1.37	80.67 ± 3.64	48.92 ± 0.52	77.31 ± 3.86
1.3B FWE EN FW2 FR 90%	33.79 ± 1.38	82.35 ± 3.51	53.21 ± 0.52	77.31 ± 3.86
2.7B FWE EN RPJ2 FR	31.82 ± 1.36	80.67 ± 3.64	51.31 ± 0.52	81.51 ± 3.57
2.7B FWE EN RPJ2 FR 90%	36.53 ± 1.41	84.87 ± 3.3	52.63 ± 0.52	78.99 ± 3.75
CroissantLLM	28.74 ± 1.32	78.15 ± 3.8	50.52 ± 0.52	78.15 ± 3.8
TransWebLLM	37.64 ± 1.42	67.23 ± 4.32	48.32 ± 0.52	58.82 ± 4.53
Qwen2.5 1.5B	36.44 ± 1.41	75.63 ± 3.95	49.69 ± 0.52	74.79 ± 4.0
EuroLLM 1.7B	33.7 ± 1.38	80.67 ± 3.64	51.34 ± 0.52	75.63 ± 3.95
Helium-1 2B	39.52 ± 1.43	78.15 ± 3.8	59.61 ± 0.51	79.83 ± 3.69
Bloom 3B	31.14 ± 1.35	78.99 ± 3.75	51.0 ± 0.52	79.83 ± 3.69
Qwen2.5 3B	40.03 ± 1.43	77.31 ± 3.86	58.59 ± 0.51	78.15 ± 3.8

Table 25: Evaluation of our models against public models on FrenchBench multiple choice tasks for French language. All evaluations are zero-shot.

Model Name	Include	Kaleidoscope	XNLI	French Topic NLI
1.3B FWE EN RPJ2 FR	44.15 ± 2.43	19.82 ± 1.44	46.1 ± 1.0	33.33 ± 1.93
1.3B FWE EN FW2 FR	42.0 ± 2.41	21.0 ± 1.48	46.22 ± 1.0	33.33 ± 1.93
1.3B FWE EN FWE FR	37.71 ± 2.37	21.39 ± 1.49	48.11 ± 1.0	36.17 ± 1.96
1.3B FWE EN RPJ 2 90%	45.35 ± 2.43	20.47 ± 1.46	48.84 ± 1.0	38.33 ± 1.99
1.3B FWE EN FW2 FR 90%	45.82 ± 2.44	19.82 ± 1.44	47.71 ± 1.0	33.33 ± 1.93
2.7B FWE EN RPJ2 FR	48.21 ± 2.44	20.87 ± 1.47	48.39 ± 1.0	33.33 ± 1.93
2.7B FWE EN RPJ2 FR 90%	21.52 ± 1.49	48.07 ± 1.0	33.33 ± 1.93	
CroissantLLM	41.05 ± 2.41	19.29 ± 1.43	49.32 ± 1.0	33.33 ± 1.93
TransWebLLM	39.86 ± 2.39	19.42 ± 1.43	47.27 ± 1.0	33.5 ± 1.93
Bloom 1B	36.04 ± 2.35	21.78 ± 1.5	46.71 ± 1.0	33.5 ± 1.93
Qwen2.5 1.5B	39.86 ± 2.39	23.23 ± 1.53	45.26 ± 1.0	41.83 ± 2.02
EuroLLM 1.7B	44.87 ± 2.43	20.21 ± 1.46	47.51 ± 1.0	33.33 ± 1.93
Helium-1 2B	47.49 ± 2.44	21.52 ± 1.49	50.8 ± 1.0	33.5 ± 1.93
Bloom 3B	42.24 ± 2.42	20.87 ± 1.47	47.67 ± 1.0	33.33 ± 1.93
Qwen2.5 3B	47.26 ± 2.44	25.98 ± 1.59	45.14 ± 1.0	59.5 ± 2.01

Table 26: Evaluation of our models against public models on Regional knowledge and NLI tasks for French language. All evaluations are zero-shot.