How Can Synthetic Data Improve Multilingual Language Model Pretraining? A Data Quality Perspective

Tongyao Zhu[1,2] **Qian Liu**[2] **Chang Ma**[4] **Jinghan Zhang**[5]
Longxu Dou[2] **Junxian He**[5] **Shiqi Chen**[3*]
[1]National University of Singapore [2]Sea AI Lab [3]City University of Hong Kong
[4]The University of Hong Kong [5]The Hong Kong University of Science and Technology
tongyao.zhu@u.nus.edu schen438-c@my.cityu.edu.hk

Abstract

Low-resource languages challenge multilingual LLMs due to limited high-quality training data, leading to weaker performance on complex reasoning and knowledge tasks. To address this, we propose improving training data quality through data synthesis, moving beyond simple resource scaling. First, we introduce SynTrans, which translates high-quality, knowledge-rich English data into low-resource languages during pretraining to inject world knowledge, though at the cost of semantic fluency. To overcome low-quality data issues while maintaining fluency, we also propose SynRank. SynRank leverages synthetic data as positive samples to train a classifier that ranks and filters noisy real-world data, enabling the extraction of high-quality subsets without expensive human cleaning. Experiments show SynRank matches handcrafted rule-based filtering by human experts and significantly improves knowledge-intensive task performance at the same filtering rate. Remarkably, higher filtering rates even improve performance with less data, demonstrating the efficiency and effectiveness of our method, surpassing expert filtering. Lastly, we introduce DA-QwenScore, a training-free metric that evaluates corpus quality by normalizing model loss with diversity measures, further enhancing evaluation efficiency. Our insights into knowledge injection could advance low-resource multilingual LLM development.[1]

1 Introduction

Large Language Models (LLMs) often struggle with low-resource languages, a widely recognized long-tail problem (Winata et al., 2020; Magueresse et al., 2020). Efforts to address this include expanding multilingual instruction tuning data (Alves et al., 2024; Üstün et al., 2024) and generating

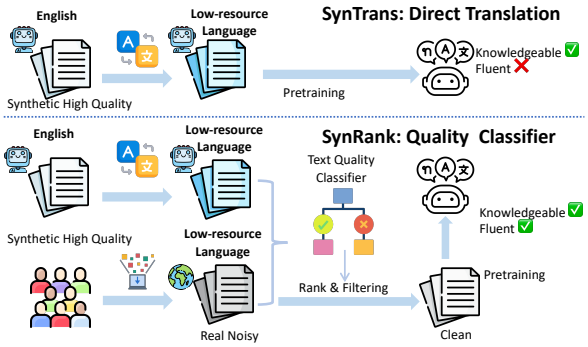

Figure 1: We propose two workflows for pretraining low-resource language models. *SynTrans* translates knowledge-rich synthetic English data into target languages, while *SynRank* cleans noisy real-world data using a quality classifier bootstrapped on synthetic and noisy samples.

synthetic parallel data (Lu et al., 2024). However, these approaches often synthesize relatively "easy" data and thus still face challenges in difficult tasks like knowledge acquisition and reasoning in low-to-middle resource languages (e.g, Chinese) (Shi et al., 2022; Huang et al., 2025). This indicates that the bottleneck in training multilingual LLMs extends beyond raw data scale, underscoring the critical role of data quality. In this work, we re-examine the problem from a data quality perspective. We identify weaknesses in multilingual pretraining data and investigate whether high-quality synthetic translations can address these gaps, as illustrated in Figure 1.

To first gain insights into the limitations of training models for low-resource languages, we evaluate several 1B-parameter LLMs trained from scratch on real-world noisy multilingual corpora under the same token budget. When comparing performance on downstream tasks across high- and low-resource languages, we observe key differences. For tasks assessing knowledge acquisition (e.g., Hellaswag, ARC-E), performance correlates linearly with language availability, which we define as the volume of the noisy corpus in the real distribution. In other words, *for languages with limited availabil-*

*Corresponding author.
[1]The code can be found at https://github.com/shiqichen17/mulsyn/tree/main

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21943–21956

ity, world knowledge within the corpora is also sparse. This indicates the primary bottleneck is not just scale, but knowledge richness, aligning with findings that a lack of specialized experts complicates high-quality data collection for low-resource languages (Doshi et al., 2024).

These initial observations motivate further exploration into leveraging English synthetic translations to incorporate more knowledge-rich multilingual data. First, we introduce *SynTrans*, which translates a high-quality English corpus (e.g., Cosmopedia) into the target language. Our exploration seeks to answer the following question: *1. Is it effective to directly use synthetic translation data for pretraining?* To investigate this, we conduct a comprehensive analysis comparing synthetic and real data across six languages, ranging from high-resource to low-resource categories. We train models from scratch and evaluate them on a range of downstream tasks, from semantic fluency to knowledge assessment. We find that while synthetic data successfully injects world knowledge, directly training on it degrades native semantic fluency.

Next, we pose another question: *2. How can we effectively utilize synthetic data without directly pretraining models on it?* To explore this, we first analyze the performance gains achieved by human expert cleaning across multiple benchmarks. Our findings reveal a strong correlation between human cleaning and improvements in challenging world-knowledge tasks. Motivated by this, we introduce *SynRank*, which uses a FastText (Joulin et al., 2017) classifier trained with synthetic data as positive samples and noisy real data as negative samples. This classifier scores and filters the noisy real data, serving as a filter to retain only high-quality documents. This solves a critical zero-resource problem: extracting high-quality native text without human-annotated positive sets. *SynRank* consistently outperforms noisy data. Moreover, unlike rule-based filtering, which produces a corpus of fixed size, *SynRank* allows for threshold adjustments to accommodate different token budgets, giving it the potential to surpass human expert rule-based filtering.

However, the aforementioned studies assess data quality by pretraining models and evaluating them on downstream tasks—an approach requiring extensive computational resources. To predict data quality without the need for full model training, we introduce a training-free metric, **DA-QwenScore**. This measures the loss of a given corpus using a reference model (*Qwen-0.5B* (Bai et al., 2023)), normalized by corpus diversity (unigram count), allowing for efficient and systematic quality estimation.

In summary, our core contributions are:

- **Knowledge vs. Fluency Analysis:** We demonstrate empirically that the primary deficit in low-resource pretraining data is complex world knowledge, rather than basic semantic fluency.
- **Low-Resource Data Filtering (*SynRank*):** We propose a reliable pipeline to bootstrap quality filters for noisy web corpora using synthetic translations, bypassing the need for human-annotated positive sets.
- **Training-Free Quality Estimation (DA-QwenScore):** We introduce a robust, diversity-adjusted perplexity metric that predicts dataset pretraining value without the computational overhead of training models.

2 Experiment Setup

Language Selection We evaluate our method in a multilingual setting, beginning with language selection. We choose Vietnamese (Vi), Thai (Th), Bahasa Indonesia (Id), Swahili (Sw), Turkish (Tr), and Chinese (Zh), as they represent a range of resource availability—from high-resource languages like Id and Tr to mid-resource languages like Vi and Zh, and lower-resource languages like Th and Sw[2] Additionally, these languages belong to distinct language families, allowing us to assess data quality across diverse linguistic contexts. Notably, low-resource languages pose significant challenges in obtaining sufficient web-based pretraining data, making them particularly relevant for evaluation.

Data Preparation Next, we prepared three types of data for our analysis and experiments. For synthetic data, we used NLLB 3.3B (Team et al., 2022) to translate the Cosmopedia (Ben Allal et al., 2024) dataset into all target languages. We selected Cosmopedia due to its rich knowledge content. For real multilingual data, we incorporated two human-collected datasets for comparison: MADLAD-400-Noisy and MADLAD-400-Clean (Kudugunta et al., 2024) (we abbreviate as MADLAD-Noisy and MADLAD-Clean in our paper). Madlad-Noisy

[2]We do not choose extremely low-resource languages because pretraining needs sufficient tokens. One such example, Wolof, can be found in Appendix C.

Language	Id	Vi	Th	Tr	Zh	Sw
Resource Availability	high	middle	low	high	middle	low
# Madlad-Noisy docs (M)	120.9	92.8	19.0	107.0	29.3	1.3
# Madlad-Clean docs (M)	38.0	55.0	17.4	56.4	19.9	0.5
Clean / Noisy rate	30%	60%	90%	50%	70%	40%
# Cosmopedia Translation Docs (M)	30.0	30.0	30.0	30.0	30.0	30.0

Table 1: Amount of data for clean and noisy sets in MADLAD-400 corpus, and amount of Cosmopedia corpus (all the amount is number of passages with the unit in millions, abbreviated as 'M'). Clean/noisy rates are rounded to the nearest 10.

consists of raw data gathered from local websites, while Madlad-Clean is a refined subset of this data, filtered and audited using handcrafted rules developed by human experts (the authors) to ensure high-quality content. The data volume and filtering rate vary across languages, as shown in Table 1.

Model and Evaluation To compare corpus quality, we pretrain 1.1B TinyLLaMA (Zhang et al., 2024) parameter models from scratch using 20B unique tokens for all languages, except for Swahili (Sw), where the available corpus was limited to 10B unique tokens. Our evaluation framework consists of four multilingual benchmarks: XCOPA (Ponti et al., 2020), XARC-Easy (Clark et al., 2018), XARC-Challenge (Clark et al., 2018) (abbreviated as XARC-E and XARC-C), and XHellaswag (Zellers et al., 2019), covering all target languages. We adopt XARC-C, XARC-E, and XHellaswag from Dac Lai et al. (2023); however, not all our chosen languages are included in Dac Lai et al. (2023). For those missing (Th, Sw, Tr, Zh), we follow the authors to leverage ChatGPT to generate translations. We will open-source these translated benchmarks to support further research. These four datasets cover a diverse range of tasks: from basic world-knowledge reasoning (XARC-E) to more complex reasoning (XARC-C), as well as natural language inference (XCOPA) and an interdisciplinary task combining NLI and easy world-knowledge reasoning (XHellaswag). Examples of the corresponding English datasets are provided in Table 6. Using the same benchmark across languages allows direct comparison. We utilize all evaluations in zero-shot setting.

3 Challenges in Limited Resources

Finding: Models pretrained on low-resource language corpora lack general knowledge, rather than semantic fluency.

First, we begin with the question: *What is the primary challenge faced by low-resource languages in pretraining?* Specifically, does the challenge stem from poor performance on certain types of downstream tasks? To investigate this, we evaluate the quality of real web data across different languages. We utilize MADLAD-noisy to pretrain TinyLlama models from scratch, as it best represents the raw, real-world distribution of each language. Crucially, all models are trained on the exact same token budget (20 billion tokens) to isolate the effect of data quality from data quantity.

We first define *language availability* based on the total volume of the noisy MADLAD corpus for each language. A language is considered more available if its total noisy corpus is larger; otherwise, it is deemed less available. We then investigate which benchmarks are most affected by these differences in availability.

As shown in Figure 2, performance on XARC-E and XHellaswag exhibits a nearly linear relationship with language availability, with correlation coefficients of 0.85 and 0.92, respectively. Notably, this strong relationship holds despite all models sharing the same compute budget. This suggests that XARC-E and XHellaswag, which involve relatively simple world-knowledge, show clear performance gains strictly when the underlying language has broader real-world resource availability.

In contrast, this correlation does not hold for XARC-C and XCOPA. For XARC-C, the lack of correlation is likely due to the highly complex world-knowledge reasoning required, which may not be present even in larger noisy corpora. For XCOPA, which serves as our proxy for native semantic fluency, the flat scaling suggests that basic linguistic fluency does not automatically improve just by having a larger overall noisy corpus. While these initial trends establish a baseline regarding the knowledge deficit in low-resource languages, we will rigorously isolate and contrast how synthetic data impacts semantic fluency versus complex world knowledge in Sections 4 and 5.

Overall, these results highlight a strong link between the richness of easy world knowledge in a corpus and the availability of language resources in real-world settings: languages with broader usage tend to have more comprehensive and informative corpora. Consequently, *low-resource languages often lack sufficient easy world knowledge*, reinforcing the need to deliberately enhance the knowledge content of pretraining corpora for these languages.

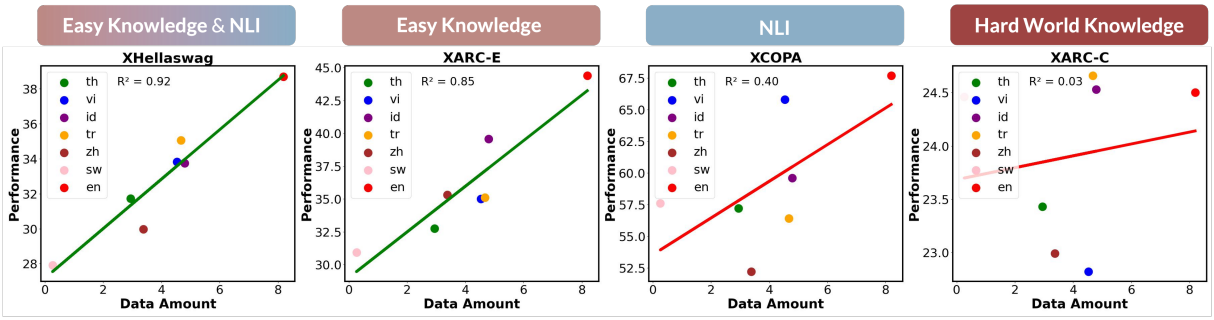

Figure 2: We examine the correlation between performance and the amount of noisy data, which serves as a proxy for data availability in real-world scenarios, across four benchmarks. The results show a strong correlation between data availability and benchmark performance on XHellaswag and XARC-E (correlation greater than 0.5), while the correlation for XCOPA and XARC-C is comparatively weaker (less than 0.5). The x-axis is log-scaled (e.g., 4 represents 2^4 million documents, etc.).

Data Type	NLI		Knowledge		Interdispline
	COPA	XNLI	ARC-E	ARC-C	HellaSwag
	Acc	Acc	Acc	Acc	Acc
Noisy	67.7	42.9	44.4	24.5	38.7
Clean	**66.7**	**46.7**	47.8	25.4	**42.1**
Cosmopedia	62.0	41.9	**49.8**	**28.6**	39.4

Table 2: Results for English data across different benchmarks, including all metrics for each benchmark. Results are shown in 10^{-2}. We can observe that Cosmopedia is rich in world knowledge.

4 Is Synthetic Translation Data Enough for Pretraining?

Finding: Pretraining with translated synthetic data enhances world knowledge but does not necessarily improve fluency.

Based on the previous finding, we know that low-resource languages struggle with world knowledge tasks, prompting a natural question: *can we find a suitable English corpus for translation to specifically address the shortcomings of low-resource languages?* In this section, we propose using translated English synthetic data for pretraining as a potential solution.

4.1 Identifying a Knowledge-Intensive English Corpus

To address the knowledge gap, we seek to identify an English corpus that provides strong world knowledge for translation. We evaluate three English corpora: MADLAD-noisy, MADLAD-clean, and Cosmopedia (Ben Allal et al., 2024) by training a 1B model on 20B tokens sampled from each corpus. The model's performance is assessed on several English benchmarks: COPA and XNLI for natural language inference (NLI), ARC-C and ARC-E for world-knowledge reasoning, and Hellaswag, which we consider an interdisciplinary benchmark combining NLI and world knowledge due to its similarity to story completion tasks requiring contextual and factual understanding. Table 6 provides examples from these four benchmarks.

Table 2 presents the performance comparison between synthetic and real data across various tasks. In ARC-Challenge, synthetic data outperforms MADLAD-400-clean by 3.2 absolute points and MADLAD-400-noisy by 4.1 points. Similarly, in ARC-Easy, synthetic data surpasses clean data by 2.0 points and noisy data by 5.4 points. For Hellaswag, synthetic data outperforms noisy data by 0.7 points but remains below clean data. However, for NLI tasks (COPA and XNLI), synthetic data underperforms compared to both noisy and clean real data. These results suggest that while synthetic data lacks fluency and depth in semantic understanding, it effectively captures a broad range of world knowledge.

4.2 SynTrans: Synthetic Translation as Pretraining Data

We propose SynTrans, an intuitive approach to augmenting data in low-resource languages. SynTrans involves translating a high-resource language corpus (e.g., English) into the target low-resource language, enabling direct use of the translated corpus for pretraining. To evaluate this approach, we use NLLB to translate the English Cosmopedia dataset into multiple languages. The results in Table 3, show that the performance of translated synthetic data follows a similar pattern to that of the original synthetic English data.

For the natural language inference (NLI) task XCOPA (Ponti et al., 2020), translated data performs worse than noisy real data in all languages

except Chinese. Closer examination reveals that this anomaly is due to the extremely low quality of the MADLAD Chinese corpus, which results in significantly lower XCOPA scores for real data. This suggests that while translated data can be beneficial, it faces limitations in capturing natural language inference ability and semantic fluency. Furthermore, across all languages, translated data achieves a consistent XCOPA score of around 55, indicating that variations in translation quality have little impact on the model's NLI capabilities.

In contrast, for world-knowledge reasoning tasks (XARC-C and XARC-E), direct translation performs well. In XARC-C, 5 out of 6 languages achieve the highest scores, while in XARC-E, 3 out of 6 languages achieve the highest scores. As for the average score across six languages, SynTrans achieves the highest score for both XARC tasks. It demonstrates that *despite its limitations in semantic tasks, direct translation effectively preserves a significant amount of world knowledge, making it valuable for commonsense reasoning tasks*.

5 SynRank: Pretraining Data Cleaning with Quality Ranker

Findings:
1. Human expert cleaning of pretraining data preserves more world knowledge.
2. Training a text classifier on synthetic data to rank and clean real data automatically improves world knowledge without sacrificing fluency.

While SynTrans provides a straightforward method to enhance world knowledge in low-resource languages through synthetic translation, it has inherent limitations. Machine translation quality varies across languages, and translated data often lacks local knowledge while struggling with semantic fluency in natural language inference tasks. This motivates the need for an approach that preserves the natural properties of real data while enriching world knowledge.

A key challenge in low-resource languages, as shown in Table 1, is data quality—many languages have a poor clean-to-noisy ratio, highlighting the prevalence of low-quality real data. This raises an important question: How do human experts improve data quality, and can we replicate this process automatically? To investigate this, we first analyze expert rule-based cleaning, discovering that human filtering primarily enhances challenging world knowledge (e.g., science and technology) rather than simply improving fluency or easy world-knowledge reasoning. Building on this insight, we propose SynRank, a classifier-based method designed to automatically filter and rank real data, retaining high-quality documents without requiring manual annotation.

5.1 Effects of Expert-Rule Cleaning

As expert filtering plays a crucial role in improving data quality, we ask: What specific aspects of the data do human experts prioritize during cleaning? Can we quantify their impact on model performance? To investigate this, we compare the performance improvement from noisy to clean datasets across all languages on XARC-E, XARC-C, XHellaswag, and XCOPA. The results, presented in Figure 3 and detailed in Table 3, reveal that expert cleaning consistently improves performance on XARC-C across all languages. However, for XARC-E and XCOPA, the improvement is inconsistent. This suggests that human experts do not primarily filter data based on fluency or simple world knowledge. Instead, they focus on preserving more complex world knowledge, such as scientific and technological content.

5.2 SynRank: Classifier-based Filtering

Inspired by the application of text classifiers for data quality filtering, we apply a classifier on multilingual datasets with the goal of automatically selecting knowledge-rich and high-quality documents from a noisy collection. Given a noisy corpus $\mathcal{D}_{noisy}$ and a high-quality corpus $\mathcal{D}_{high}$, we train a text classifier $f(t)$ that learns to decide whether a piece of text t belongs to $\mathcal{D}_{noisy}$ or $\mathcal{D}_{high}$. The classifier f assigns a score to any text t: $f(t) = P(t \in \mathcal{D}_{high})$.

A naive approach for obtaining $\mathcal{D}_{high}$ would be to manually select documents from $\mathcal{D}_{noisy}$ or use widely recognized high-quality corpora, as is common in high-resource languages (e.g., selecting high-quality English corpora (Li et al., 2024)). However, this approach is largely infeasible for low-resource languages due to the lack of human expert availability for annotation and the absence of labeled high-quality corpora. To overcome these challenges, we automate the process by translating a high-quality English corpus into a low-resource language, which then serves as $\mathcal{D}_{high}$.

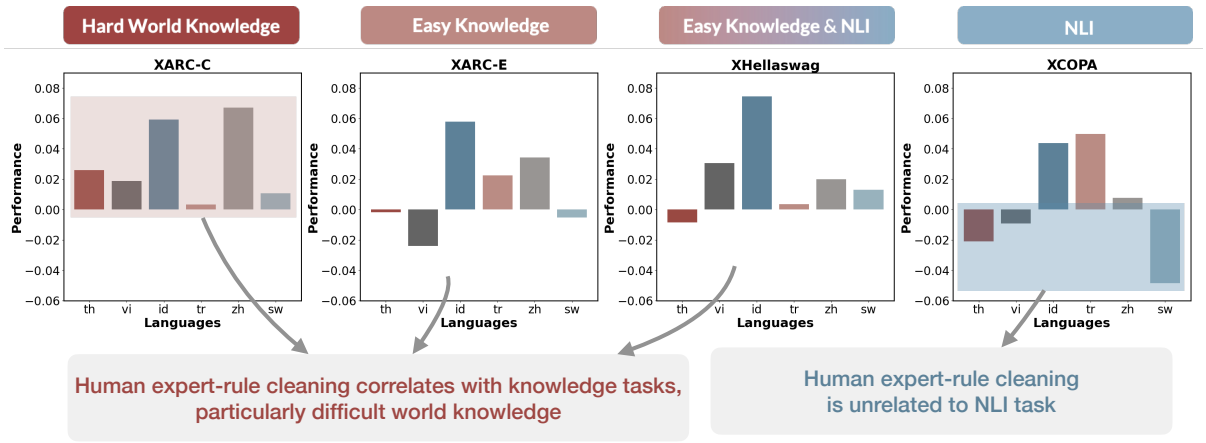

Figure 3: The correlation between performance improvements of models trained on noisy versus clean corpora in Madlad-400 across all languages is examined on four benchmarks. We observe a consistent performance gain in XARC-C for all languages. For XARC-E and XHellaswag, most languages show performance improvements. However, for XCOPA, half of the languages demonstrate an increase, while the other half do not. These findings suggest that human-expert rule cleaning is particularly beneficial for knowledge acquisition tasks, especially for challenging knowledge, but is unrelated to NLI tasks.

Benchmark	Method	Language						
		Thai	Vietnamese	Indonesian	Turkish	Chinese	Swahili	Average
XCOPA	Madlad-Noisy	**57.20**	**65.80**	59.60	56.40	52.20	**57.60**	58.13
	Madlad-Clean	56.00 ↓1.20	65.20 ↓0.60	62.20 ↑2.60	**59.20** ↑2.80	52.60 ↑0.40	54.80 ↓2.80	**59.67** ↑1.54
	SynRank	56.40 ↓0.80	63.40 ↓2.40	**62.40** ↑2.80	59.00 ↑2.60	50.20 ↓2.00	57.00 ↓0.60	58.07 ↓0.06
	SynTrans	54.00 ↓3.20	58.00 ↓7.80	56.40 ↓3.20	54.20 ↓2.20	**55.60** ↑3.40	57.40 ↓0.20	55.93 ↓2.20
XHellaswag	Madlad-Noisy	**31.72**	33.82	33.74	35.06	29.97	27.91	32.04
	Madlad-Clean	31.45 ↓0.27	**34.85** ↑1.03	**36.25** ↑2.51	35.18 ↑0.12	30.57 ↑0.60	28.27 ↑0.36	**32.76** ↑0.72
	SynRank	31.36 ↓0.36	34.35 ↑0.53	34.16 ↑0.42	**34.83** ↑0.77	29.79 ↑0.17	28.21 ↑0.30	32.12 ↑0.08
	SynTrans	29.99 ↓1.73	33.74 ↓0.08	35.07 ↑1.33	35.14 ↑0.08	**31.45** ↑1.48	**29.42** ↑1.51	32.47 ↑0.43
XARC-E	Madlad-Noisy	32.76	35.02	39.57	35.12	35.31	30.92	34.78
	Madlad-Clean	32.70 ↓0.06	34.18 ↓0.84	**41.86** ↑2.29	35.91 ↑0.79	36.52 ↑1.21	30.76 ↓0.16	35.32 ↑0.54
	SynRank	**33.54** ↑0.78	35.52 ↑0.50	40.79 ↑1.22	36.14 ↑1.02	35.71 ↑0.40	30.18 ↓0.74	35.31 ↑0.53
	SynTrans	32.52 ↓0.24	**35.73** ↑0.71	39.79 ↑0.22	**37.26** ↑2.14	**39.53** ↑4.22	**31.03** ↑0.11	**35.98** ↑1.20
XARC-C	Madlad-Noisy	23.43	22.82	24.53	24.66	22.99	24.46	23.82
	Madlad-Clean	24.04 ↑0.61	23.25 ↑0.43	25.98 ↑1.45	24.74 ↑0.08	24.53 ↑1.54	24.72 ↑0.26	24.54 ↑0.72
	SynRank	23.51 ↑0.08	22.91 ↑0.09	26.67 ↑2.14	25.17 ↑0.51	23.42 ↑0.43	**25.67** ↑1.21	24.56 ↑0.74
	SynTrans	**24.30** ↑0.87	**26.67** ↑3.85	**27.26** ↑2.73	**25.17** ↑0.51	**27.26** ↑4.27	25.41 ↑0.95	**26.01** ↑2.19

Table 3: Results for models trained on various corpora, including noisy and clean sets from Madlad, our SynTrans corpus, and our SynRank corpus, evaluated across four benchmarks in Thai, Vietnamese, Indonesian, Turkish, Chinese, and Swahili. The results also include overall scores, which are the averages of the four benchmarks. For fair comparison for SynRank and human experts cleaning, we control the same clean/noisy rate for all the languages.

In our experiments, we demonstrate that translating a knowledge-intensive corpus, such as Cosmopedia, effectively creates $\mathcal{D}_{high}$ for selecting high-quality documents in low-resource languages. This approach aligns with the idea of leveraging synthetic data as a taxonomy of world knowledge. The overall process is illustrated in Figure 1.

Implementation and Efficiency To construct the training set for the classifier, we randomly sample 10,000 documents from the translated Cosmopedia as positive examples and 10,000 documents from the noisy web corpus as negative examples. For each document, we truncate the text to the first 512 tokens, as the beginning of a document typically contains sufficient salient information for quality assessment. Following Occam's razor, we employ a FastText classifier (Joulin et al., 2017) rather than more computationally expensive multilingual Transformer encoders. We initialize the model using the input embeddings of Qwen2.5-0.5B as pretrained vectors. The model is trained for 5 epochs with a learning rate of 0.1, an embedding

Language	Precision	Recall	F1
Indonesian (id)	99.50	99.80	99.65
Swahili (sw)	98.75	99.82	99.28
Turkish (tr)	98.98	99.75	99.36
Vietnamese (vi)	99.52	99.74	99.63

Table 4: Intrinsic evaluation of the FastText quality ranker on a held-out test set comprising 10,000 synthetic Cosmopedia documents (positive) and 10,000 noisy MADLAD documents (negative).

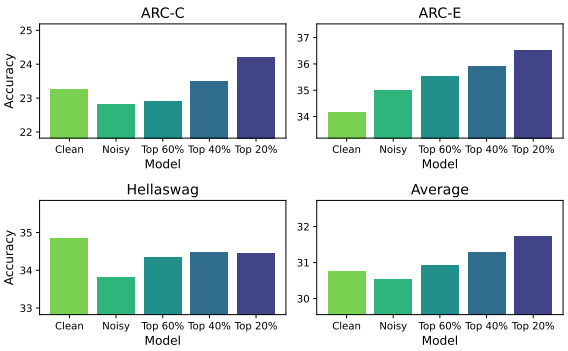

Figure 4: Performance comparison of different thresholds of filtering with SynRank on Vietnamese. The evaluations are done in zero-shot. We exclude XCOPA as even expert-rule cleaning decreases performance.

dimension of 896, up to 3-grams (`wordNgrams=3`), and a minimum word frequency count of 3.

This design choice prioritizes high scalability. Running inference with our FastText classifier achieves a throughput of approximately 980 documents per second on a standard CPU. This makes the pipeline highly parallelizable and practical for scoring massive web-scale corpora, whereas running heavy multilingual encoders would require extensive GPU resources with limited room for inference optimization.

Intrinsic Evaluation To ensure the reliability of the quality ranker, we intrinsically evaluate its classification performance on a completely held-out test set consisting of 10,000 positive and 10,000 negative documents. As shown in Table 4, the classifier achieves near-perfect F1 scores across the tested languages. This exceptionally high accuracy indicates that the synthetic translated Cosmopedia possesses a distinctly different, cleaner distribution compared to the noisy web documents, allowing a lightweight classifier to easily distinguish high-quality, knowledge-dense text from web noise.

Data Filtering To train the SynRank classifier, we sample 10,000 positive documents from the translated Cosmopedia corpus and 10,000 negative documents from the noisy CommonCrawl web corpus for each language. To avoid overfitting to the specific stylistic patterns of the synthetic Cosmopedia data and obtain a smoother distribution, we adopt an iterative training approach. After training the initial classifier, we apply it to score the CommonCrawl corpus. We then select the highest-scoring real documents and add them to the positive training set for a second round of training. This refined positive set—now containing both synthetic knowledge-rich data and high-quality native text—helps the classifier better maintain native semantic fluency.

Finally, we apply this first-iteration trained classifier to rank and filter the entirety of the original noisy corpus, retaining only the top-scored documents. To ensure a fair comparison in our downstream pretraining experiments, we apply the exact same filtering rate as the clean-to-noisy ratio of the original MADLAD corpus, maintaining an equivalent data budget across datasets.

As shown in Table 3, applying SynRank filtering successfully extracts high-quality content, leading to performance gains over noisy corpora. The SynRank corpus outperforms the noisy dataset in 5 out of 6 languages on easy world knowledge acquiring tasks XHellaswag and XARC-E. For the more complex world knowledge task, XARC-C, all languages benefit from filtering. In the NLI task of XCOPA, the average XCOPA score only drops by 0.06 absolute points. Given that XCOPA contains only 500 samples and exhibits substantial variability in performance, we can conclude that SynRank does not significantly harm XCOPA performance compared to the noisy corpus.

To conclude, the SynRank corpus improves performance on world knowledge tasks without compromising fluency, demonstrating the effectiveness of classification-based filtering. When considering average performance across all benchmarks, the advantages of the filtered data become clear. Notably, when filtering the same proportion of data, this method achieves results comparable to human-expert-filtered corpora, highlighting its potential as an efficient, scalable alternative for improving low-resource language data quality.

Scaling via Filtering Rate To evaluate the impact of the filtering threshold, we vary the data selection threshold and observe its effect on model performance. Figure 4 presents the results for Vietnamese, where we observe a clear trend: as the selection threshold increases, the quality of

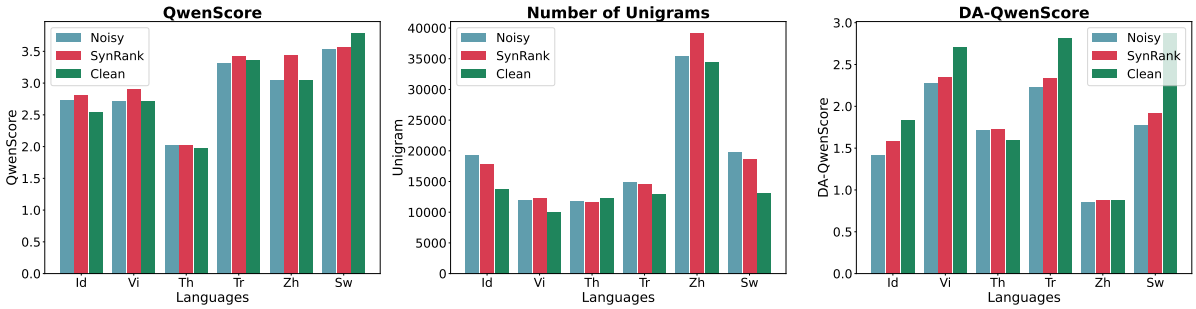

Figure 5: QwenScore, Unigram, and DA-QwenScore for all languages from left to right. For DA-QwenScore, a clear pattern emerges: clean data performs better than SynRank data, which in turn outperforms noisy data, consistent with the overall downstream performance.

the filtered corpus improves, leading to steady performance gains on XARC and XHellaswag. This finding suggests that our method provides a practical mechanism for balancing cost and performance. When data is abundant and computational resources are sufficient, selecting only the top-ranked documents ensures a cleaner, higher-quality corpus. Conversely, when resources are constrained, a moderate threshold can be used to maintain reasonable data quality while reducing filtering intensity.

6 Measuring Data Quality Heuristically

Our earlier experiments demonstrate that both SynRank filtering and manual cleaning significantly alter the quality of pretraining data, leading to varying model performance. However, downstream evaluation requires substantial computational resources, as each model must be fully pretrained before testing. This raises an important question: Can we estimate pretraining data quality and predict model performance before training begins?

To effectively evaluate data quality before training, we first use a straightforward heuristic: assessing the corpus using a well-trained reference model, Qwen-0.5B (Bai et al., 2023). We calculate the average loss on the given corpus and define it as **QwenScore**. Denoting a data sample as $x^{(i)} = \{w_1^{(i)}, w_2^{(i)}, \ldots, w_{T_i}^{(i)}\}$, where $w_t^{(i)}$ represents tokens in a document and T_i represents the sequence length of the i-th sample from a corpus $\mathcal{D}$, the average loss is calculated as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\log P(w_t^{(i)}|w_1^{(i)}, \ldots, w_{t-1}^{(i)}) \quad (1)$$

The underlying assumption is that when evaluating a corpus using a high-quality model, a higher loss reflects that the data contains more complex, knowledge-dense structures that are challenging to learn. However, computing the raw loss across MADLAD-noisy, MADLAD-clean, and SynRank datasets yields inconsistent rankings (Figure 5, left). This occurs because raw loss is highly sensitive to out-of-distribution noise; unstructured web junk (e.g., typos, random URLs, scraped boilerplate) artificially inflates the vocabulary size and produces high loss values that do not correlate with actual linguistic quality.

Diversity-Adjusted QwenScore (DA-QwenScore). To isolate true textual complexity and penalize noisy datasets driven by unstructured web junk, we propose normalizing the loss against corpus diversity. The **DA-QwenScore** is defined as $\mathcal{L}_{\mathcal{N}} = \mathcal{L}/N_{uni}$, where N_{uni} is the number of unique unigrams in the corpus. By dividing the loss by the unigram count, we heavily penalize noisy corpora whose high loss is merely a byproduct of inflated vocabularies. As shown in Figure 5, the DA-QwenScore ranking is largely consistent across all six languages (*Clean* > *SynRank* > *Noisy*), aligning exactly with the downstream performance of our trained models on knowledge-intensive benchmarks (Table 3).

Metric Robustness and Generalization. To ensure that DA-QwenScore is not overfitted to our specific evaluator model or diversity measure, we conducted ablations using an alternative model (Llama-3.2-1B) and an alternative diversity metric (unique bigrams). As shown in Figure 6, the clear and consistent quality ranking holds regardless of the reference model or n-gram order used. This confirms that the diversity-adjusted metric is a robust, training-free proxy for evaluating dataset quality.

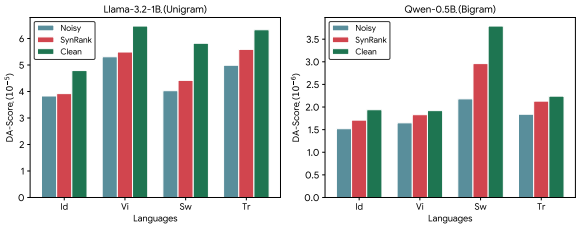

Figure 6: Robustness ablations for the DA-Score metric. The consistent ranking (*Clean* > *SynRank* > *Noisy*) holds when using a different evaluator model (Llama-3.2-1B, scaled by 10^{-5}, using unigrams) or a different diversity measure (Qwen-0.5B, using bigrams, scaled by 10^{-6}).

7 Related Work

Synthetic Data for Improving Multilingual LLMs. Synthetic data has been an effective approach to overcoming data scarcity. Existing efforts include including synthetic data during the supervised fine-tuning stage (Taori et al., 2023; Wang et al., 2023), as well as during the pretraining stage for better knowledge acquisition (Ben Allal et al., 2024; Liu et al., 2024). Another line of work focuses on scaling up multilingual training with data synthesis. As translators like NLLB (Team et al., 2022), Google Translator (Johnson et al., 2017) exhibits superior performance in multilingual translation compared to LLM, we could use the translators to create parallel translation data. This approach has been widely used to scale multilingual LLM to hundreds of languages (Alves et al., 2024; Lu et al., 2024) and boosts performances on low-resource languages (Üstün et al., 2024; Liu et al., 2024). Other work directly uses self-instruct to curate synthetic data (Wei et al., 2023; Guo et al., 2024). However, most of study in this domain focus on only translation ability (Wang et al., 2025), which is straightforward using parallel data. Our study, instead, also explores the effects of more challenging knowledge and reasoning tasks, which highlight the pros and cons of synthetic pretraining data across languages.

Pretraining Data Selection. The selection of pretraining data remains a critical challenge at scale. While simple heuristics like deduplication (Lee et al., 2022; Tirumala et al., 2023) and rule-based filtering (Longpre et al., 2024) are effective for basic cleaning, more nuanced selection often relies on classifier-based approaches (Brown et al., 2020; Xie et al., 2023; Li et al., 2024; Weber et al., 2024). These methods train a text classifier using high-quality documents as positive examples to rank and filter the broader corpus. However, these prior works depend heavily on the availability of human-curated positive English documents, making them difficult to apply to low-resource languages. Our work extends the classifier-based approach to this data-scarce regime in multilingual setups. By leveraging synthetic translated data as a proxy positive signal, we introduce a reliable pipeline to bootstrap quality filters for zero-resource languages where human-annotated positive sets simply do not exist.

8 Conclusion

We investigate the role of synthetic data in pretraining language models for low-resource languages. Our findings indicate that the primary challenge these languages face is a lack of world knowledge. To address this issue, we propose two methods: *SynTrans* (direct translation) and *SynRank* (classifier-filtered noisy-to-clean data refinement) for pretraining. Notably, our results show that *SynRank* enhances world knowledge without compromising fluency. We hope this research inspires future work on improving data quality and availability for low-resource languages. In particular, further investigation is needed to assess the effects of integrating synthetic and real data, as well as leveraging multiple languages.

Limitations

While our study provides a robust pipeline for low-resource languages, several limitations remain. First, due to computational constraints, we evaluate 1B-parameter models using a 20B token budget and single-source, monolingual pretraining. Future work must investigate how SynRank scales to larger architectures, substantially larger pretraining budgets, and mixed multilingual corpora evaluated on advanced reasoning tasks.

Second, our methodology contains an inherent English-centric bias. Translating English corpora injects factual knowledge but risks transferring Western cultural norms, highlighting the need to balance synthetic data with culturally representative native text. This bias extends to our evaluation framework; translated benchmarks (like XHellaswag and XARC) often miss language-specific nuances. Consequently, there is a pressing need for the community to develop native, non-English-derived benchmarks to accurately evaluate lower-resourced languages.

Ethical Considerations

Our work uses open-source corpora to pretrain language models and also uses an LLM for automatic translation. While there might be bias in these corpora and LLM-generated data, we do not think it poses significant ethical concerns or any potential risks beyond research purposes.

Acknowledgments

We thank the anonymous reviewers and the Area Chairs for their constructive feedback, which significantly improved this manuscript. Tongyao Zhu is funded by the EDB-IPP program with Sea AI Lab.

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. Cosmopedia.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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:1803.05457v1*.

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

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Do not worry if you do not have data: Building pretrained language models using translationese. *Preprint*, arXiv:2403.13638.

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

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and He-Yan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *arXiv preprint arXiv:2502.07346*.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

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

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, and 40 others.

2024. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data. *Preprint*, arXiv:2404.07503.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages. *arXiv preprint arXiv:2407.05975*.

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

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

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. *Journal of machine learning research*, 21(140):1–67.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *Preprint*, arXiv:2210.03057.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. 2023. D4: Improving llm pretraining via document de-duplication and diversification. *Advances in Neural Information Processing Systems*, 36:53983–53995.

Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, David Ifeoluwa Adelani, Yihong Chen, Raphael Tang, and Pontus Stenetorp. 2025. Multilingual language model pretraining using machine-translated data. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28087–28107, Suzhou, China. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. *Preprint*, arXiv:2411.12372.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui,

Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model. *Preprint*, arXiv:2307.06018.

Genta Indra Winata, Guangsen Wang, Caiming Xiong, and Steven Hoi. 2020. Adapt-and-adjust: Overcoming the long-tail problem of multilingual speech recognition. *arXiv preprint arXiv:2012.01687*.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems (NeurIPS)*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *Preprint*, arXiv:2402.07827.

A Pretraining Details

We adopt the open-source TinyLlama codebase to train the 1B LLMs. For All six languages except Swahili (we use 10B tokens to train Swahili models cause the MADLAD-clean and MADLAD-noisy only contain 10B unique tokens), we sample 20B unique tokens from the corpus. Following the original training recipe, we use a cosine learning rate schedule, with a max learning rate of 4e-4, and a minimum learning rate of 4e-5, and warmup the model for 2000 steps. We use a global batch size of 512, with a sequence length of 2048. The optimizer is AdamW, and we enable gradient clipping of 1. The model architecture exactly follows the original TinyLlama model[3] with around 1.1B parameters.

B Classifier Training Details

For the classifier used in SynRank, we use 10,000 positive and 10,000 negative documents as the input to `fasttext`. The positive documents are randomly sampled from the translated Cosmopedia corpus, and the negative documents are sampled from the CommonCrawl corpus in each language. To avoid overfitting the Cosmopedia corpus, we adopt an iterative approach: after training a classifier for the first round, we add the top-scored documents from CommonCrawl into the positive document set. Therefore, in the second round, the positive set includes both translated Cosmopedia, as well as high-quality CommonCrawl documents. We verify that this iterative approach achieves better performance in maintaining semantic fluency. We use the classifier after the first iteration in all experiments to score and filter a corpus.

[3]https://huggingface.co/TinyLlama/TinyLlama-1.1B-step-50K-105b

C Additional Experiments

To further evaluate the limitations and capabilities of our approach in data-constrained environments, we extend our experiments to include Wolof. We classify Wolof as an "extremely low-resource" language regarding available web corpora for pretraining; specifically, the Madlad noisy dataset contains only approximately 93 million unique tokens for this language.

We conducted pretraining of 1B-scale language models from scratch using this limited data. We evaluated performance using the AfriXNLI and AfriMMLU benchmarks. The results are presented in Table 5. Under such severe token constraints (training on roughly 100M to 500M tokens), the models largely achieve random baseline performance. For example, the 33.3% score on AfriXNLI corresponds to random guessing for a 3-class classification task. However, we observe that the SynTrans method remains comparable to, or slightly better than, the noisy baseline on the AfriMMLU benchmark, suggesting potential resilience even in extremely low-data regimes.

D Comparing English Corpora

We extend our study of the relationship between DA-QwenScore and data quality to common English corpora. We follow the experiment setting in Penedo et al. (2024) and compare seven corpus: C4, dolma, redpajama, pile, Refinedweb, Fineweb, SlimPajama (Raffel et al., 2020; Soldaini et al., 2024; Weber et al., 2024; Gao et al., 2020; Penedo et al., 2024; Soboleva et al., 2023). Penedo et al. (2024) show that these seven types of corpus have a distinguishable performance on downstream tasks. We show the QwenScore and NA-QwenScore's ranking in Figure 7.

Table 5: Performance comparison on Wolof benchmarks. We report accuracy scores on AfriXNLI and AfriMMLU. Note that a score of 33.3 on AfriXNLI represents random baseline performance.

Model	**Tokens**	**AfriXNLI** (Native Direct)	**AfriMMLU** (Direct)	**AfriMMLU** (Translate)
Madlad-noisy	100M	33.3	20.8	20.8
Madlad-noisy	500M	33.3	22.6	19.8
SynTrans	**500M**	**33.3**	**22.6**	**21.6**

Figure 7: QwenScore and DA-QwenScore on seven english corpus.

E Evaluation Examples

We show sample instances of our evaluation benchmarks in Table 6.

F Licenses of Artifacts

The MADLAD-400 corpus is released with the CC-BY-4.0 license. The TinyLlama library uses Apache-2.0 license. We follow the guidance to use the artifacts for research purpose only.

G Use of LLMs

LLMs, including ChatGPT and Gemini, are used for polishing the text and fixing grammar errors. They were not used for ideation or implementation.

Benchmark	Type	Question	Choice 1	Choice 2	Choice 3	Choice 4
ARC-E	Easy Knowledge	Which of the following is an example of an assistive device?	contact lens	motorcycle	raincoat	coffee pot
ARC-C	Challenging Knowledge	George wants to warm his hands quickly by rubbing them. Which skin surface will produce the most heat?	dry palms	wet palms	palms covered with oil	palms covered with lotion
Hellaswag	Semantic/ Easy Knowledge	A female chef in white uniform shows a stack of baking pans in a large kitchen presenting them. the pans	contain egg yolks and baking soda.	are then sprinkled with brown sugar.	are placed in a strainer on the counter.	are filled with pastries and loaded into the oven.
COPA	Semantic	My body cast a shadow over the grass.	The grass was cut.	The sun was rising.	-	-

Table 6: Examples of the four benchmarks we use in our evaluation. For better understanding, we use the English examples. In evaluation, the benchmarks are translated into the target language.