TOWARD CROSS-LINGUAL QUALITY CLASSIFIERS FOR MULTILINGUAL PRETRAINING DATA SELECTION

Yassine Turki, Vinko Sabolčec, Bettina Messmer, & Martin Jaggi
Machine Learning Optimization Lab
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Lausanne, Switzerland
{yassine.turki,vinko.sabolcec,bettina.messmer,martin.jaggi}@epfl.ch

ABSTRACT

As Large Language Models (LLMs) scale, data curation has shifted from maximizing volume to optimizing the signal-to-noise ratio by performing quality filtering. However, for many languages, native high-quality data is insufficient to train robust quality classifiers. This work investigates the idea that quality markers in embedding space may show cross-lingual consistency, which would allow high-resource languages to subsidize the filtering of low-resource ones. We evaluate various filtering strategies, including cross-lingual transfer, third quartile sampling (Q3), and retention rate tuning. Our results demonstrate that massive multilingual pooling frequently outperforms monolingual baselines in both rank stability and aggregate accuracy for a 1B model trained on 103B tokens, delivering gains for high resource languages (1.2% increase in aggregate normalized accuracy for French) and matching or exceeding monolingual baselines for low-resource languages. However, we find that scale alone does not guarantee stability. Furthermore, for high-resource languages like French, we show that refining the decision boundary through third quartile sampling (Q3) or tuning the retention rate is necessary to fully leverage the multilingual signal.

1 INTRODUCTION

As machine learning architectures and optimization strategies have matured, the focus of the community has shifted toward the foundational element of model performance: data quality. Previous work has demonstrated that "more" is not synonymous with "better" (Raffel et al., 2020; Penedo et al., 2023). Building on this insight, recent efforts have shown that aggressively curating high-quality subsets from web-scale corpora can match or exceed the performance of models trained on far larger datasets (Li et al., 2025; Penedo et al., 2024; Messmer et al., 2025).

Historically, data curation relied on heuristic-based filtering (Nguyen et al., 2023; Raffel et al., 2023; Rae et al., 2022; Penedo et al., 2023). However, the emergence of model-based classifiers has enabled a more nuanced selection of semantically dense content. FineWeb-Edu (Penedo et al., 2024) reached the performance of LLMs trained on $350B$ tokens of raw English data using only 10% of tokens through LLM-based filtering. Building upon this, FineWeb2-HQ (Messmer et al., 2025) extended model-based filtering to multilingual data, training classifiers for 20 languages and demonstrating that 15% of tokens could match the performance of models trained on the full FineWeb2 (Penedo et al., 2025) dataset.

While model-based filtering has proven effective for high-resource languages, the multilingual domain faces a critical challenge: many languages lack sufficient native high-quality data to train effective standalone classifiers. This work investigates whether quality classifiers can generalize across languages by exploiting shared semantic structures in multilingual embedding spaces, which would enable high-resource languages to effectively support filtering for low-resource ones.

We hypothesize that quality is a measurable density of information and logical coherence rather than subjective preference. We assume high-quality text may be distinguished in the embedding space by **grammatical coherence**, where formal structures create distinct activation patterns; **lexical density**, characterized by specialized vocabulary rather than boilerplate; and **information density**, derived

from high logical and factual coherence. If this hypothesis holds, a classifier $\mathcal{C}$ trained to recognize quality patterns in French should transfer to typologically distant languages like Chinese, either because they share an underlying quality manifold, or because structural and formatting regularities in the positive anchor datasets (such as Wikipedia markup or instruction-tuning templates) are themselves cross-lingual proxies for quality. Our primary contributions are as follows:

- **Multilingual Synergy:** We demonstrate that massive multilingual pooling frequently outperforms monolingual baselines in both rank stability and aggregate accuracy, delivering gains for high-resource languages and beating state-of-the-art baselines for low-resource ones.
- **Empirical Validation of Cross-Family Transfer:** We demonstrate that classifiers trained on one language family can effectively curate high-quality tokens in typologically distant languages.
- **Q3 Sampling Strategy:** We introduce a third-quartile (Q3) sampling strategy that trains against fluent but low-utility text. We find this refinement sharpens decision boundaries and improves performance in high-resource languages such as French and Spanish.

2 Related Work

Evolution of Web Data Curation. Early large-scale datasets relied primarily on heuristic-based filtering of Common Crawl. Wenzek et al. (2019) introduced CCNet, which utilized FastText (Joulin et al., 2016) for language identification and perplexity-based scoring. This approach was refined by Raffel et al. (2020) with C4 and Penedo et al. (2023) with RefinedWeb, the latter demonstrating that aggressive deduplication and string-matching heuristics could allow web data to match curated dataset performance. However, Penedo et al. (2024) showed that simple heuristics fail to capture the nuanced educational value required for complex reasoning tasks.

Cross-lingual Representation Learning Modern NLP has moved beyond language-specific models toward unified multilingual encoders. Models such as XLM-RoBERTa (Conneau et al., 2020) utilize a Transformer architecture trained on a masked language modeling objective across 100+ languages simultaneously. The core power of these models lies in their ability to align semantic clusters across languages within a shared embedding space. During pre-training, the model learns that semantically equivalent words occupy similar topological positions regardless of surface form. For example, the English word "Science" and the German word "Wissenschaft" are positioned near each other in the 768-dimensional embedding space, as both relate similarly to concepts like "logic" and "fact." This alignment enables classification based on geometric position in latent space rather than surface-level syntax.

Model-Based Quality Filtering. A paradigm shift occurred with model-based classifiers. FineWeb-Edu (Penedo et al., 2024) used Llama-3 (Grattafiori et al., 2024) as a judge to score educational quality and create a knowledge-rich dataset. To scale this approach, Li et al. (2025) and Messmer et al. (2025) used lightweight classifiers based on FastText (Joulin et al., 2016) or MLPs trained on embeddings of an XLM-RoBERTa model. Our work builds directly upon the FineWeb2-HQ pipeline by Messmer et al. (2025).

Multilingual Scaling. While English-centric curation is well-established, multilingual curation presents unique challenges. Nguyen et al. (2023), Kudugunta et al. (2023) and Penedo et al. (2025) expanded web-scale cleaning to hundreds of languages using perplexity and basic heuristics. FineWeb2-HQ (Messmer et al., 2025) advanced this by applying model-based filtering to 20+ languages. Despite these advances, two areas remain underexplored. First, while multilingual encoders are widely used, the degree to which one classifier can generalize across language families (e.g., from Nordic to Romance) has not been systematically characterized. Second, prior work primarily uses random negative sampling; the potential of smarter sampling techniques to refine decision boundaries remains largely uninvestigated. Our work addresses both gaps.

3 METHODS

3.1 CLASSIFIER DATASETS

To train a robust multilingual quality classifier, we curate a diverse set of high-quality "positive" samples and contrast them against a baseline of general web data. Our strategy extends the FineWeb2-HQ framework by expanding their high-quality anchors (**MKC+**) with additional instruction-based and synthetic sources to form the **MKC-e** (Extended) dataset.

Positive Anchors (MKC+). We incorporate the original FineWeb2-HQ anchors, which prioritize structured, knowledge-dense content. These datasets are Multilingual MMLU (OpenAI, 2024), Aya Dataset and Collection (Singh et al., 2024), OpenAssistant-2 (Köpf et al., 2023) and Include-Base-44 (Romanou et al., 2024).

Extended Positives (MKC-e). To generalize the classifier's ability to recognize natural queries and encyclopedic prose, we expand the MKC+ pool with:

- **Tagengo (Devine, 2024):** A multilingual chat dataset containing approximately 75,000 conversations in 74 languages between humans and GPT4 (OpenAI et al., 2024).
- **MURI-IT (Wikipedia Subset) (Köksal et al., 2024):** A dataset containing instruction-output pairs across 200 languages. We specifically extracted the **Wikipedia subset** to ensure samples reflect factual, encyclopedic prose and are knowledge-rich. Furthermore, MURI-IT contains samples combining different languages. As our goal is to ablate classifier behaviour for different languages, we decided not to include multiple languages in a given sample.
- **EuroBlocks-SFT-Synthetic (Martins et al., 2025):** Multilingual synthetic data for Supervised Fine-Tuning used to train the EuroLLM 9B Instruct model. It spans 35 languages.
- **WikiQA (Apertus et al., 2025)**: A dataset linking real-world user queries to factual Wikipedia answer sentences. With 65 languages, it provides a stronger focus on low-resource languages.

Negative Anchors (FineWeb2). We use the raw FineWeb2 (Penedo et al., 2025) corpus as our source of negative samples. We assume a random sample from this web-scale crawl primarily contains boilerplate, informal prose, or noise.

3.2 SAMPLING AND DATA PREPARATION

Balancing and Preprocessing. To ensure consistency, all samples are processed into 768-dimensional embeddings using the XLM-RoBERTa encoder (Conneau et al., 2020) used in the original FineWeb2-HQ release. Consistent with prior work, we perform minimal preprocessing (concatenation of prompt/response) and remove samples with `<unk>` tokens.

For training, we sample $100,000$ positive documents per language, an increase from the $80,000$ used in (Messmer et al., 2025) to improve representation. To counter class imbalance in low-resource languages, we upsample positives by a maximum factor of 3 (analogous to (Chung et al., 2023)), ensuring high-resource languages do not dominate the gradient without overfitting to duplicated samples. We sample an equal number of negative documents to maintain a balanced class distribution. For our classifier, we use a simple MLP, with a single hidden layer (256 dim, ReLU, 20% dropout) and sigmoid output, trained to predict positive and negative classes based on XLM-RoBERTa embeddings in the same way as FineWeb2-HQ.

Negative Sampling Strategies. We employ two distinct strategies for selecting negative samples from FineWeb2:

1. **Random Sampling:** The standard approach, selecting documents uniformly at random to distinguish quality content from general web noise.

2. **Q3 (Hard Negatives):** To sharpen the decision boundary, we sample documents that score in the 50th–75th percentile (the third quartile) of a preliminary classifier. These "Q3 negatives" typically represent fluent but low-utility text (e.g., repetitive procedural content), forcing the model to learn subtler distinctions beyond surface-level fluency.

4 EVALUATION

Qualitative Analysis. To understand how each classifier filters samples, we first verify that score distributions concentrate near zero with a flatter tail at higher scores. We then inspect the top and bottom 25 samples to confirm that highly-ranked samples are knowledge-rich and well-structured, while low-ranked ones are poorly written or uninformative. Using a monolingual FineWeb2-HQ classifier as baseline, we compute rank correlations (Spearman and Kendall) and analyze the top 20 samples with the largest rank increases and decreases across classifiers. This reveals the filtering patterns each classifier prioritizes (e.g., grammar, content quality, symbols). Selected ranking changes are shown in Appendix C.

Downstream Task Evaluation. To better understand how different filtering methods would influence the performance of an LLM, we conduct experiments by training a small 1B parameter Apertus (Apertus et al., 2025) architecture model on the filtered FineWeb2 by a given classifier. The model is trained on 103B tokens with a sequence length of 4096, and sees each token at most twice, except for Arabic, where for retention rates of 10% and 20%, we have replicated the filtered data 10x and 5x respectively, to account for a lower number of tokens. These tokens are directly from the filtered samples of the classifier, followed by a rehydration step as described in (Penedo et al., 2025). Technical details for the LLM training can be found in Appendix E.

To evaluate the models, we use the Language Model Evaluation Harness library (Gao et al., 2024). Our main criterion is normalized accuracy, as recommended by Kydlíček et al.. The tasks we use span multiple capabilities, including knowledge retrieval, reasoning and natural language understanding. Full benchmark list can be found in Appendix G. The average rank across these benchmarks serves as our final measure of a filtering strategy's robustness. We also report the mean normalized accuracy, as the average rank can be volatile for very small differences in performance.

5 EXPERIMENTS AND RESULTS

We conduct ablations across four languages representing diverse linguistic profiles: French (Romance, high-resource), Spanish (Romance, high-resource), Arabic (Semitic, morphologically complex), and Chinese (Sino-Tibetan, logographic). Our experimental design addresses three core questions:

1. **Multilingual Synergy:** Does pooling data from multiple languages improve performance compared to monolingual baselines?
2. **Cross-lingual Transfer:** Can classifiers trained on typologically distant languages identify quality in a target language?
3. **Decision Boundary Refinement:** Can the Q3 strategy improve classifier precision in high-resource settings?

Baseline Configurations. We establish two primary baselines: **No filtering**, representing random sampling from FineWeb2 (lower bound), and **HQ**, a monolingual classifier trained following the FineWeb2-HQ pipeline (Messmer et al., 2025) with MKC+ anchors.

Multilingual Synergy. We first investigate whether data quality is a general feature shared across samples from different languages, or if it is language-specific. Our hypothesis is that there exist shared regularities in XLM-RoBERTa's embedding space that correlate with quality. If we train a model to detect these regularities across multiple languages, then we would obtain a general classifier that could recognize high-quality samples in unseen languages, and even boost the data quality for languages it was trained on. In order to test this hypothesis, we train a general multilingual classifier (denoted by **ML**). We use the **MKC-e** pool to account for the lack of samples for some

Table 1: Comparison of models trained without filtering, with monolingual high-quality (HQ) filtering, and with our multilingual (ML) classifier on Chinese. The ML strategy achieves the highest aggregate accuracy and best average rank, supporting the hypothesis that quality features are transferable.

Benchmark	No filtering	HQ	ML
Agieval Cn	0.3618	**0.3644**	0.3457
ARC	0.2855	0.3145	**0.3171**
Belebele_c	0.3011	0.3200	**0.3222**
Ceval-valid	0.2288	0.2489	**0.2615**
Cmmlu_c	0.3206	0.3471	**0.3608**
GMMLU_c	0.2800	0.3075	**0.3200**
Include_c	0.3468	0.3523	**0.3541**
MMLU_c	0.2772	0.2940	**0.2987**
PAWS	0.5520	0.5535	**0.5610**
Xcopa	0.5860	0.5920	**0.6080**
XNLI	0.3546	0.4072	**0.4189**
Xstorycloze	0.6559	**0.6625**	0.6625
XWinograd	0.6806	**0.6825**	0.6766
Aggregate acc_norm	0.4024	0.4190	**0.4236**
Average rank	2.85	1.77	**1.31**

Table 2: Comparison of models trained without filtering, with monolingual high-quality (HQ) filtering, and with our multilingual (ML) classifier on Spanish. The ML strategy achieves the highest aggregate accuracy and best average rank, supporting the hypothesis that quality features are transferable.

Benchmark	No filtering	HQ	ML
ARC-Challenge	0.2991	0.3077	**0.3248**
Belebele_c	0.3422	**0.3533**	0.3456
GMMLU_c	0.3100	0.3150	**0.3250**
HellaSwag	0.5006	0.5263	**0.5310**
Include_c	0.3491	0.3727	**0.3891**
M_MMLU_c	0.2814	0.2985	**0.3050**
XNLI	0.4618	0.4538	**0.4783**
Aggregate acc_norm	0.3635	0.3753	**0.3855**
Average rank	2.86	2.00	**1.14**

knowledge, and also to balance the dominance of the Aya Collection dataset in terms of size and representation. The detailed counts for each language can be found in the appendix A.

Tables 1 and 2 evaluate the effect of multilingual quality filtering on downstream performance for a 1B model in Chinese and Spanish. The multilingual classifier (ML) achieves the highest aggregate accuracy and lowest average rank in both languages, outperforming both No filtering and monolingual high-quality filtering (HQ).

Improvements are observed consistently across diverse reasoning and natural language understanding benchmarks, such as ARC Clark et al. (2018), GMMLU Singh et al. (2025), XNLI (Conneau et al., 2018), and Include (Romanou et al., 2024). While HQ occasionally yields marginal gains on individual tasks, ML provides more stable and stronger overall performance.

These results are consistent with the hypothesis that data quality corresponds to a shared structure in XLM-RoBERTa's representation space. However, we cannot determine whether this reflects abstract semantic features or cross-lingual formatting regularities in the positive anchor datasets. By jointly modeling quality across languages, the classifier captures features that transfer effectively across linguistic boundaries, leading to improved downstream generalization.

Table 3: Spearman and Kendall correlations between family-specific classifiers and the French HQ baseline. High correlation in distant families (e.g., Nordic) indicates a shared quality manifold, while the drop in "Romance (no French)" suggests potential syntactic interference.

Experiment	Spearman	Kendall
Romance (spa, fra, por, ita, ron, cat) MKC+	0.8928	0.7173
Nordic (swe, dan, nob, isl) MKC+	0.8820	0.6990
Romance, no french (spa, por, ita, ron, cat) MKC-e	0.7139	0.5228

Cross-lingual Transfer. We have seen that multilingual transfer is possible since a classifier trained on multiple languages performs better than its monolingual counterpart. We note that our ML classifier was trained on these languages; therefore, it already has knowledge in that language. We hypothesize that this knowledge has been augmented by the data from the other languages. Can this ML classifier actually generalize to unseen languages? Would this classifier be able to perform as well (if not better) on a language like French if it has never seen a sample in that language?

To answer these questions, we investigate the performance of a classifier on French. First, we train classifiers on different language families (i.e. Romance, Nordic, Germanic, etc.) and apply them on the French split of FineWeb2 to obtain scores. Then, we compute the Spearman and Kendall correlations between the scores of these classifiers and the HQ baseline. We would expect to have a very strong correlation score in the Romance family (French, Spanish, Portuguese, Italian, Romanian, Catalan) and a low correlation score in the other language families, which are linguistically distant from French. We display some of these correlation scores in Table 3, and provide the full table for other language families in the Appendix C.1. We can see that intra-family transfer is strong: Romance (MKC+) $\rho_s = 0.89$, indicating near-identical document rankings despite training on multiple Romance languages. This is to be expected, as languages in the same family share the same root and possess a big overlap in vocabulary. What is more interesting is that the classifiers trained on unrelated families, for example Nordic (MKC+): $\rho_s = 0.88$, have nearly matching Romance correlation despite limited lexical and syntactic overlap. Finally, we can see that removing French from Romance hurts more than expected. Romance without French (MKC-e) drops to $\rho_s = 0.71$, correlating worse than some linguistically distant families like Germanic or Uralic. This suggests potential syntactic interference: when the classifier trains on closely related but distinct languages (Spanish, Italian, Portuguese), it may learn Romance-specific syntactic patterns that don't perfectly generalize to French, obscuring the underlying quality signal.

To investigate this further, we plot the score distribution of the Nordic classifier in Figure 1 to give us insights on how the classifier behaves. We apply it on a held-out set of French high-quality samples, negative samples, and the general FineWeb2. As we can observe, the classifier recognizes almost perfectly the negative and positive classes. It applies a very strict threshold (90th percentile: 0.027) and assigns generally lower scores, yet still identifies high-quality content effectively. Additional experiments, including the baseline plot, can be found in the Appendix B.

Additionally, we train our 1B models with the datasets from our classifiers. Despite the high rank correlation, we would expect the classifier trained on Romance to perform better than Nordic, even though both have never seen French samples in their training data. And we expect the Romance classifier to perform close to the HQ baseline, as these languages should be similar enough to French to give the classifier a good idea of what a high-quality sample should be.

We show the results of the experiment in Table 4, which provides a surprising result. The Romance classifier has the same rank as the HQ baseline, however, with slightly less in normalized accuracy. Furthermore, the Nordic classifier is the one with the highest mean normalized accuracy, which even outperformed the HQ baseline trained on French.

The success of the Nordic classifier provides evidence for transferable quality features across language families. Because Nordic languages (Swedish, Danish, Norwegian, Icelandic) share no lexical or syntactic overlap with French, the classifier cannot rely on surface-level patterns.

These results suggest that cross-lingual transfer of quality signals is practically effective even across distant language families, though whether this reflects abstract semantic structure or shared format-

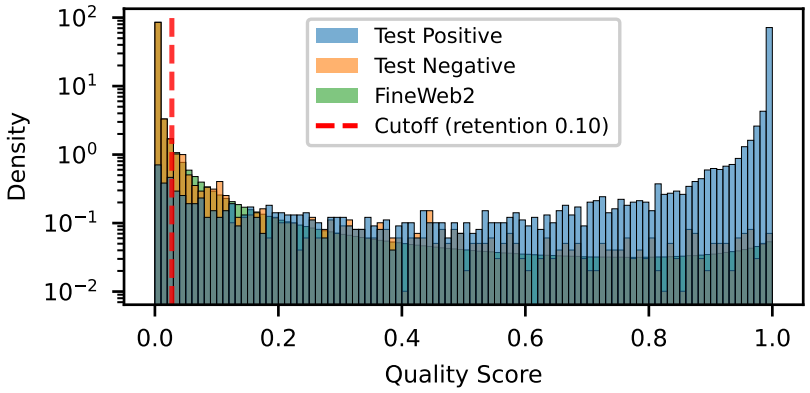

Figure 1: French quality score distribution from a Nordic classifier. The classifier effectively separates high-quality (positives) from low-quality (negatives) French text despite being trained solely on Nordic languages, suggesting the practical effectiveness of cross-lingual quality transfer.

Table 4: Models trained on French data filtered by "Nordic" and "Romance (no French)" classifiers. The Nordic classifier outperforms the native French HQ baseline in aggregate accuracy, providing evidence for transferable quality signals across language families.

Benchmark	No filtering	HQ	Romance (no fra)	Nordic
ARC-Challenge	0.2891	0.3071	0.3054	**0.3157**
Belebele_c	0.3444	0.3511	**0.3689**	0.3422
GMMLU_c	0.2625	0.2925	0.2750	**0.3075**
HellaSwag	0.4883	0.4748	**0.4908**	0.4673
Include_c	0.3866	0.4153	0.4272	**0.4535**
M_MMLU_c	0.2831	0.2945	**0.2950**	0.2936
XNLI	0.4707	**0.4855**	0.4827	0.4783
XWinograd	0.6386	0.6506	0.6024	**0.6627**
Aggregate acc_norm	0.3954	0.4089	0.4059	**0.4151**
Average rank	3.50	**2.12**	**2.12**	2.25

ting regularities in the embedding space remains to be characterized. However, we note that these high rank correlations may partially reflect the classifier learning shared dataset artifacts such as Wikipedia formatting or instruction-tuning templates rather than a purely abstract representation of quality. Nonetheless, capturing these cross-lingual artifacts serves as a highly effective and robust proxy for pretraining data selection. Conversely, the underperformance of Romance-without-French suggests that syntactic similarity can be a double-edged sword: the classifier may overfit to Spanish/Italian grammatical structures that don't perfectly align with French, creating systematic biases that distant language families avoid.

Refining the Decision Boundary. We have observed that training on multiple languages provides strong performance boosts, and also can generalize to unseen samples. Nevertheless, upon examining the training pipeline, one could argue that using random samples from FineWeb2 might not be the best strategy for selecting negative samples. Standard negative sampling draws randomly from FineWeb2, teaching the classifier to distinguish quality content from typical web noise (advertisements, navigation menus, broken formatting, spam). However, as filtering becomes more precise, a subtler distinction emerges: *separating high-quality content from fluent but low-utility text*.

The Q3 strategy addresses this by sampling negatives from the 50th-75th percentile of the baseline HQ distribution. These samples are grammatically correct and well-formatted, but lacking in information density, logical depth, or educational value. We provide an example in the Appendix D.

Table 5: Comparison of benchmark results using standard random negatives vs. Q3 negatives (fluent but low-utility text) in French. Q3 sampling consistently yields superior aggregate performance, with the monolingual classifier (Q3) performing slightly better than the ML (Q3) classifier.

Benchmark	No filtering	HQ	Q3	ML	ML (Q3)
ARC-Challenge	0.2891	0.3071	0.3105	**0.3165**	0.3054
Belebele_c	0.3444	0.3511	**0.3711**	0.3511	0.3633
GMMLU_c	0.2625	0.2925	0.3025	0.2900	**0.3075**
HellaSwag	0.4883	0.4748	0.4727	**0.4956**	0.4911
Include_c	0.3866	0.4153	0.4272	0.4224	**0.4582**
M_MMLU_c	0.2831	0.2945	0.2992	0.2927	**0.3007**
XNLI	0.4707	0.4855	**0.4876**	0.4807	0.4767
XWinograd	0.6386	0.6506	**0.6627**	0.6024	0.6145
Aggregate acc_norm	0.3954	0.4089	**0.4167**	0.4064	0.4147
Average rank	4.50	3.00	**2.00**	3.00	2.38

Table 6: Comparison of benchmark results using standard random negatives vs. Q3 negatives (fluent but low-utility text) in Spanish. Q3 sampling consistently yields superior aggregate performance.

Benchmark	No filtering	HQ	ML	ML (Q3)
ARC-Challenge	0.2991	0.3077	0.3248	**0.3282**
Belebele_c	0.3422	**0.3533**	0.3456	**0.3533**
GMMLU_c	0.3100	0.3150	0.3250	**0.3275**
HellaSwag	0.5006	0.5263	0.5310	**0.5372**
Include_c	0.3491	0.3727	0.3891	**0.4000**
M_MMLU_c	0.2814	0.2985	0.3050	**0.3083**
XNLI	0.4618	0.4538	**0.4783**	0.4667
Aggregate acc_norm	0.3635	0.3753	0.3855	**0.3887**
Average rank	3.86	2.86	2.00	**1.14**

In this approach, once a classifier is obtained, we apply it on FineWeb2 and extract samples from the third quartile (i.e. 50th to 75th percentile) as our negative samples for the training of a new classifier. We conduct these experiments using the HQ baseline (Q3) and the ML classifier (ML Q3). The ML classifier uses negative samples from the third quartile of its scores. We apply this extra step to all languages that have more than 200,000 samples in FineWeb2 to ensure we will not need aggressive token replication.

Tables 5 and 6 evaluate the impact of refining the negative sampling strategy using third-quartile bootstrapping (Q3). Compared to random negatives from FineWeb2, harder negatives consistently improve performance, confirming that sharper decision boundaries emerge when the classifier is trained on more ambiguous examples.

For French, the monolingual Q3 classifier achieves the strongest aggregate performance, while the multilingual bootstrapped model (ML Q3) remains highly competitive, with only a marginal drop in normalized accuracy. In contrast, for Spanish, ML Q3 yields the best overall performance across benchmarks, surpassing both standard multilingual filtering and monolingual HQ.

Taken together, these results indicate that bootstrapped negative sampling systematically strengthens quality discrimination. While language-specific refinement can yield peak in-language performance, the multilingual Q3 model matches or exceeds these gains in some languages and, critically, preserves cross-lingual transfer without retraining. This further supports the existence of shared in representation space that can be progressively refined through Q3 sampling.

Tuning the Retention Rate. While FineWeb2-HQ was derived using a retention rate of 10% for high-resource languages, we argue that this hyperparameter should be tuned for each language separately. The rate for which we filter samples for a language can have a big impact on the filtered data.

Table 7: Comparison between 10% and 15% retention using the ML classifier. Increasing the retention rate to 15% improves accuracy across most benchmarks, suggesting the default 10% threshold is overly aggressive when using our classifier on French.

Benchmark	ML	ML (15%)
ARC-Challenge	**0.3165**	0.3139
Belebele_c	0.3511	**0.3611**
GMMLU_c	**0.2900**	0.2875
HellaSwag	0.4956	**0.5135**
Include_c	0.4224	**0.4726**
M_MMLU_c	0.2927	**0.2991**
XNLI	**0.4807**	**0.4807**
XWinograd	0.6024	**0.6386**
Aggregate acc_norm	0.4064	**0.4209**
Average rank	1.62	**1.25**

Table 8: Comparison between 10% and 15% retention using the ML classifier on Spanish. Similar to French, the 15% retention rate yields a higher aggregate accuracy and better average rank when using our ML classifier.

Benchmark	ML	ML (15%)
ARC-Challenge	0.3248	**0.3436**
Belebele_c	0.3456	**0.3500**
GMMLU_c	**0.3250**	0.3225
HellaSwag	0.5310	**0.5440**
Include_c	0.3891	**0.4036**
M_MMLU_c	0.3050	**0.3080**
XNLI	**0.4783**	0.4562
Aggregate acc_norm	0.3855	**0.3897**
Average rank	1.71	**1.29**

Tables 7, 8, and 9 analyze the effect of tuning the retention rate used during multilingual and monolingual filtering. Increasing the retention rate consistently improves aggregate performance in French and Spanish, with ML at 15% outperforming the default 10% across most benchmarks. Gains are particularly pronounced on knowledge-intensive tasks such as Include (Romanou et al., 2024) and HellaSwag (Zellers et al., 2019), indicating that overly aggressive filtering can discard useful high-quality content.

In Arabic, instead of using 56% like for FineWeb2-HQ, we try a retention rate of 10% and 20%, resulting in a replication of tokens of 10x and 5x, respectively. We see that the standard retention of 56% yields the best overall performance when combined with multilingual filtering. This suggests that optimal retention rates are language-dependent, and that aggressive repetition of the same high-quality data does not necessarily lead to the best performance. We further note that the ML gain over HQ for Arabic (0.5%) is comparable to the seed variance reported in Appendix H (0.3%), and should therefore be interpreted with caution pending multi-seed validation.

Synthesis: When Does Multilingual Pooling Help? Across our four languages, we observe a pattern: the key insight is not that multilingual pooling *always* dominates, but that it provides a reliable baseline across languages, while language-specific optimization (negative sampling strategy, retention rate, anchor curation) can yield comparable or superior results when tuned appropriately. These findings provide evidence that the classifier learns language-agnostic markers in the embedding space by training on multiple languages. Note on Stochasticity: While multilingual pooling improves overall rank stability, our ablations show that downstream aggregate accuracy remains sensitive to the classifier's initial sampling seed (shifting by 0.3% in Arabic and 0.8% in French). We provide a more detailed analysis of these seed variances in Appendix H.

Table 9: Evaluation of standard (56%) vs. aggressive (10%, 20%) filtering with replication. Unlike high-resource languages, reducing Arabic retention to increase token replication degrades performance.

Benchmark	HQ	ML	HQ (10%)	HQ (20%)	ML (10%)	ML (20%)
ARC-Easy	**0.2898**	0.2855	0.2720	0.2771	0.2741	0.2762
AlGhafa PIQA-MT	**0.5145**	0.5112	0.5019	0.5090	0.4948	0.5128
AlGhafa RACE	0.2775	**0.2883**	0.2747	0.2834	0.2765	0.2786
AlGhafa SciQ	0.4432	0.4503	0.4503	0.4171	**0.4573**	0.4563
ARC-Challenge	0.2660	**0.2797**	0.2669	0.2772	0.2712	0.2626
Belebele_c	0.3233	0.3122	**0.3378**	0.2944	0.3278	0.3256
GMMLU_c	0.2575	0.2700	0.2475	0.2550	**0.2725**	0.2525
HellaSwag	0.3873	**0.3925**	0.3592	0.3728	0.3747	0.3857
Include_c	0.2681	**0.3043**	0.2717	0.2772	0.2844	0.2790
M_MMLU_c	0.2614	0.2646	0.2652	**0.2674**	0.2618	0.2634
AlGhafa PIQA	0.6088	0.6110	0.5958	0.6023	0.6039	**0.6143**
XNLI	0.3309	0.3349	0.3325	0.3333	0.3325	**0.3357**
XStoryCloze	**0.5923**	0.5811	0.5877	0.5817	0.5784	0.5804
Aggregate acc_norm	0.3708	**0.3758**	0.3664	0.3652	0.3700	0.3710
Average rank	3.62	**2.31**	4.31	3.69	3.69	3.23

Table 10: Aggregated macro and micro metrics for all strategies using a 10% retention rate. The ML (Q3) strategy achieves the best overall performance in both rank and normalized accuracy, highlighting the robustness of combining both our methods to obtain a robust general multilingual classifier.

Method	Macro Rank	Micro Rank	Macro Acc	Micro Acc
ML (Q3)	**1.6983**	**1.7857**	**0.4082**	**0.4113**
ML	1.9808	1.9286	0.4052	0.4092
HQ	2.4556	2.4286	0.4011	0.4052
No filtering	3.7248	3.7143	0.3871	0.3907

6 Conclusion

This work investigates whether quality classifiers can generalize across languages by exploiting shared semantic structures in multilingual embedding spaces. Through systematic evaluation across French, Spanish, Arabic, and Chinese, we demonstrate that classifiers trained on typologically distant language families can effectively filter quality content in unrelated languages. Across the four tested languages, multilingual classifiers improved over monolingual baselines, with particularly strong gains in Spanish (3.0 ranks) and Arabic (2.31 ranks). We also introduce the Q3 sampling strategy, which refines decision boundaries by training against fluent but low-utility text rather than random negatives, offering a complementary approach to multilingual pooling for high-resource languages. No single strategy dominates. For French, Q3 negatives, ML (15%), and ML (Q3) all achieve comparable results, suggesting practitioners should select based on computational constraints and available data. The 10% threshold from prior work may be overly aggressive when using our ML classifier; increasing to 15% improved French performance substantially (40.64% to 42.09%), which indicates that this hyperparameter needs language-specific tuning. However, when comparing all approaches (Table 10), we find that the ML (Q3) strategy is the one dominating in terms of performance, which means combining our methods leads to better results. Overall, our findings suggest that multilingual pooling can help democratize high-quality data curation for underrepresented languages while boosting performance on high-resource languages, though the effectiveness varies by language family and resource level.

REFERENCES

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, Julien Launay, and Badreddine Noune. AlGhafa evaluation benchmark for Arabic language models. In Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, and Rawan Almatham (eds.), *Proceedings of ArabicNLP 2023*, pp. 244–275, Singapore (Hybrid), December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.arabicnlp-1.21. URL https://aclanthology.org/2023.arabicnlp-1.21.

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, Inés Altemir Mariñas, Mohammad Hossein Amani, Matin Ansaripour, Ilia Badanin, Harold Benoit, Emanuela Boros, Nicholas Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille Challier, Clement Charmillot, Jonathan Coles, Jan Deriu, Arnout Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fan, Silin Gao, Miguel Gila, María Grandury, Diba Hashemi, Alexander Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Frederike Lübeck, Roman Machacek, Theofilos Manitaras, Andreas Marfurt, Kyle Matoba, Simon Matrenok, Henrique Mendonça, Fawzi Roberto Mohamed, Syrielle Montariol, Luca Mouchel, Sven Najem-Meyer, Jingwei Ni, Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo Paoletti, Marco Passerini, Ivan Pavlov, Auguste Poiroux, Kaustubh Ponkshe, Nathan Ranchin, Javi Rando, Mathieu Sauser, Jakhongir Saydaliev, Muhammad Ali Sayfiddinov, Marian Schneider, Stefano Schuppli, Marco Scialanga, Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Sternfeld, Ayush Kumar Tarun, Paul Teiletche, Jannis Vamvas, Xiaozhe Yao, Hao Zhao, Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, David Rosenthal, Elliott Ash, Florian Tramèr, Joost VandeVondele, Livio Veraldi, Martin Rajman, Thomas Schulthess, Torsten Hoefler, Antoine Bosselut, Martin Jaggi, and Imanol Schlag. Apertus: Democratizing open and compliant llms for global language environments, 2025. URL https://arxiv.org/abs/2509.14233.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.44. URL http://dx.doi.org/10.18653/v1/2024.acl-long.44.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019. URL https://arxiv.org/abs/1911.11641.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining, 2023. URL https://arxiv.org/abs/2304.09151.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations, 2018. URL https://arxiv.org/abs/1809.05053.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020. URL https://arxiv.org/abs/1911.02116.

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

Peter Devine. Tagengo: A multilingual chat dataset, 2024. URL https://arxiv.org/abs/2405.12612.

Olga Majewska Qianchu Liu Ivan Vuli'c Edoardo M. Ponti, Goran Glava s and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. *arXiv preprint*, 2020. URL https://ducdauge.github.io/files/xcopa.pdf.

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. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein,

Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. 2021.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*, 2023.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations, 2024. URL https://arxiv.org/abs/2405.18392.

Alexandra Institute. m_mmlu (revision 18e6c8e), 2025. URL https://huggingface.co/datasets/alexandrainst/m_mmlu.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification, 2016. URL https://arxiv.org/abs/1607.01759.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. Madlad-400: A multilingual and document-level large audited dataset, 2023. URL https://arxiv.org/abs/2309.04662.

Hynek Kydlíček, Guilherme Penedo, Clémentine Fourier, Nathan Habib, and Thomas Wolf. Finetasks: Finding signal in a haystack of 200+ multilingual tasks. URL https://huggingface.co/spaces/HuggingFaceFW/blogpost-fine-tasks.

Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, and Hinrich Schütze. Muri: High-quality instruction tuning datasets for low-resource languages via reverse instructions, 2024. URL https://arxiv.org/abs/2409.12958.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023. URL https://arxiv.org/abs/2304.07327.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese, 2024. URL https://arxiv.org/abs/2306.09212.

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, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models, 2025. URL https://arxiv.org/abs/2406.11794.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668, 2021. URL https://arxiv.org/abs/2112.10668.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 158–167, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1015. URL https://aclanthology.org/P17-1015.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *International Joint Conference on Artificial Intelligence*, 2020.

Pedro Henrique Martins, João Alves, Patrick Fernandes, , Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. Eurollm-9b: Technical report, 2025.

Bettina Messmer, Vinko Sabolčec, and Martin Jaggi. Enhancing multilingual llm pretraining with model-based data selection, 2025. URL https://arxiv.org/abs/2502.10361.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning, 2022.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages, 2023. URL https://arxiv.org/abs/2309.09400.

OALL. Alghafa arabic llm benchmark translated. https://huggingface.co/datasets/OALL/AlGhafa-Arabic-LLM-Benchmark-Translated, 2023.

OpenAI. Mmmlu dataset, 2024. URL https://huggingface.co/datasets/openai/MMMLU. Hugging Face.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor,

Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Matteo Pagliardini, Pierre Ablin, and David Grangier. The ademamix optimizer: Better, faster, older, 2024. URL https://arxiv.org/abs/2409.03137.

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

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

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. Fineweb2: One pipeline to scale them all – adapting pre-training data processing to every language, 2025. URL https://arxiv.org/abs/2506.20920.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis insights from training gopher, 2022. URL https://arxiv.org/abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL https://arxiv.org/abs/1910.10683.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011. URL https://people.ict.usc.edu/~gordon/publications/AAAI-SPRING11A.PDF.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas,

et al. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*, 2024.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2020. URL https://arxiv.org/abs/1909.08053.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning, 2024.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation, 2025. URL https://arxiv.org/abs/2412.03304.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Alexey Tikhonov and Max Ryabinin. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning, 2021.

Siyuan Wang, Zhongkun Liu, Wanjun Zhong, Ming Zhou, Zhongyu Wei, Zhumin Chen, and Nan Duan. From lsat: The progress and challenges of complex reasoning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2201–2216, 2021.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data, 2019. URL https://arxiv.org/abs/1911.00359.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. Paws-x: A cross-lingual adversarial dataset for paraphrase identification, 2019. URL https://arxiv.org/abs/1908.11828.

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

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. Jec-qa: A legal-domain question answering dataset. In *Proceedings of AAAI*, 2020.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models, 2023.

A Dataset Counts

Table 11: Dataset Counts by Language (sorted by total count, descending). Note: OA = openassistant2, MMLU = openai_mmlu, Inc = include, Tag = tagengo, Euro = euroblocks, Wiki = muri_wikipedia, AH = aya_human, AC = aya_collection, WQA = wikiqa.

Language	OA	MMLU	Inc	Tag	Euro	Wiki	AH	AC	WQA	Total
eng_Latn	61,278	99,842	0	15,771	151,135	6,657	3,944	14,693,823	0	15,032,450
jpn_Jpan	756	14,042	501	2,521	4,735	6,971	6,259	6,218,459	0	6,254,244
arb_Arab	76	14,042	552	789	0	8,508	4,995	5,857,458	17,706	5,904,126
tha_Thai	1,503	0	0	133	0	7,187	724	5,338,232	0	5,347,779
deu_Latn	5,797	14,042	139	5,739	14,081	7,009	241	4,689,989	0	4,737,037
fra_Latn	3,686	14,042	419	5,369	14,882	6,988	1,422	4,285,094	0	4,331,902

Continued on next page

Table 11 – Continued from previous page

Language	OA	MMLU	Inc	Tag	Euro	Wiki	AH	AC	WQA	Total
tel_Telu	0	0	548	0	0	7,765	8,439	4,058,535	0	4,075,287
rus_Cyrl	13,336	0	552	8,056	4,727	7,042	423	4,005,166	0	4,039,302
fin_Latn	138	0	551	92	1,022	6,970	742	3,939,941	16,383	3,965,839
spa_Latn	26,811	14,042	550	8,318	17,428	7,012	3,854	3,872,864	0	3,950,879
ita_Latn	899	14,042	548	7,063	15,963	7,100	738	3,890,852	0	3,937,205
urd_Arab	0	0	352	3	0	7,893	654	3,876,197	0	3,885,099
pol_Latn	431	0	548	1,090	5,358	7,013	1,483	3,841,451	16,964	3,874,338
por_Latn	2,581	14,042	551	12,564	13,966	7,367	8,997	3,786,062	0	3,846,130
hin_Deva	0	14,042	547	20	7,982	7,217	1,153	3,772,864	0	3,803,825
fas_Arab	0	0	548	184	0	7,504	1,578	3,785,250	5,615	3,800,679
nld_Latn	72	0	551	383	7,683	6,840	1,733	3,736,938	18,723	3,772,923
ukr_Cyrl	821	0	550	323	5,191	0	522	3,729,748	12,979	3,750,134
ces_Latn	12	0	0	179	4,105	6,793	0	3,719,214	11,203	3,741,506
heb_Hebr	24	0	550	120	0	6,916	0	3,658,066	17,229	3,682,905
cmn_Hani	0	14,042	545	5,338	27,507	9,368	4,909	3,606,935	0	3,668,644
hun_Latn	113	0	550	214	4,010	7,096	98	3,637,911	14,320	3,664,312
swe_Latn	1	0	0	256	6,476	6,524	1,310	3,632,622	14,650	3,661,839
tur_Latn	37	0	548	406	0	7,084	4,046	3,628,109	18,422	3,658,652
kor_Hang	20	14,042	500	1,609	2,905	7,343	361	3,605,894	17,616	3,650,290
cat_Latn	1,194	0	0	73	223	7,209	0	3,625,537	15,438	3,649,674
srp_Cyrl	0	0	550	6	0	7,529	152	3,636,573	0	3,644,810
ben_Beng	1	14,042	548	0	0	7,609	1,534	3,601,287	8,532	3,633,553
vie_Latn	203	0	550	429	0	7,040	8,676	3,613,270	0	3,630,168
ind_Latn	12	14,042	550	240	0	0	786	3,610,078	0	3,625,708
ron_Latn	0	0	0	71	3,993	7,170	0	3,602,212	11,188	3,624,634
bul_Cyrl	0	0	550	56	210	7,221	0	3,602,878	11,207	3,622,122
hau_Latn	0	0	0	0	0	8,045	3,512	3,608,883	0	3,620,440
tam_Taml	0	0	550	5	0	7,766	14,133	3,596,707	0	3,619,161
slk_Latn	0	0	0	17	990	7,061	0	3,594,203	15,036	3,617,307
slv_Latn	0	0	0	10	201	6,873	0	3,593,626	16,125	3,616,835
dan_Latn	40	0	0	67	4	6,348	97	3,601,900	8,212	3,616,668
ell_Grek	0	0	552	308	582	7,495	623	3,606,249	827	3,616,636
yor_Latn	0	14,042	0	0	0	0	11,758	3,587,233	0	3,613,033
zsm_Latn	0	0	501	5	0	8,060	10,073	3,593,313	0	3,611,952
bel_Cyrl	0	0	550	2	0	7,499	0	3,589,912	12,868	3,610,831
sin_Sinh	0	0	0	5	0	7,290	14,524	3,587,051	0	3,608,870
plt_Latn	0	0	0	0	0	6,895	14,597	3,586,962	0	3,608,454
ibo_Latn	0	0	0	0	0	8,767	1,534	3,597,292	0	3,607,593
swh_Latn	0	14,042	0	0	0	7,518	366	3,580,061	0	3,601,987
ary_Arab	0	0	0	0	0	0	8,090	3,591,621	0	3,599,711
glg_Latn	0	0	0	0	0	6,906	0	3,572,365	19,481	3,598,752
lit_Latn	0	0	534	11	0	7,239	916	3,573,281	16,354	3,598,335
amh_Ethi	0	0	0	3	0	7,132	1,207	3,589,993	0	3,598,335
nob_Latn	0	0	0	26	534	6,870	0	3,572,365	17,742	3,597,537
eus_Latn	257	0	500	6	0	7,075	939	3,573,304	15,069	3,597,150
ltz_Latn	0	0	0	1	0	6,999	0	3,572,365	17,689	3,597,054
som_Latn	0	0	0	0	0	7,036	7,704	3,582,111	0	3,596,851
ekk_Latn	0	0	224	18	179	7,028	0	3,572,365	16,824	3,596,638
isl_Latn	0	0	0	5	18	7,372	0	3,572,365	15,426	3,595,186
gla_Latn	0	0	0	0	0	7,655	0	3,572,365	15,053	3,595,073
mkd_Cyrl	0	0	551	5	0	7,548	0	3,572,365	14,066	3,594,535
lvs_Latn	0	0	0	22	176	7,515	0	3,572,365	14,311	3,594,389
als_Latn	0	0	551	5	0	8,152	120	3,572,485	12,705	3,594,018
ydd_Hebr	0	0	0	2	0	7,062	0	3,572,365	13,417	3,592,846
mlt_Latn	0	0	0	0	0	4,771	0	3,572,365	15,309	3,592,445
mar_Deva	0	0	0	1	0	7,743	3,545	3,579,228	0	3,590,517
cym_Latn	0	0	0	0	0	6,944	0	3,572,365	11,044	3,590,353
guj_Gujr	0	0	0	0	0	7,492	3,989	3,578,511	0	3,589,992
mal_Mlym	0	0	479	2	0	7,660	1,749	3,577,960	0	3,587,850
nno_Latn	0	0	0	0	0	0	0	3,572,365	14,518	3,586,883
npi_Deva	0	0	500	0	0	5,020	4,002	3,576,367	0	3,585,889
sna_Latn	0	0	0	0	0	7,457	1,368	3,576,309	0	3,585,134
zul_Latn	0	0	0	0	0	7,642	1,833	3,574,437	0	3,583,912
afr_Latn	0	0	0	2	0	6,503	0	3,577,285	0	3,583,790
kan_Knda	0	0	0	1	0	7,574	334	3,573,855	0	3,581,764
gle_Latn	0	0	0	0	0	6,549	1,245	3,573,610	0	3,581,404
ceb_Latn	0	0	0	0	0	7,130	727	3,573,092	0	3,580,949
mya_Mymr	0	0	0	2	0	7,367	472	3,572,837	0	3,580,678
hat_Latn	0	0	0	0	0	7,982	106	3,572,471	0	3,580,559
kaz_Cyrl	0	0	500	2	0	7,547	0	3,572,365	0	3,580,414
snd_Arab	0	0	0	0	0	7,470	274	3,572,639	0	3,580,383
azj_Latn	0	0	548	4	0	7,313	0	3,572,365	0	3,580,230
kat_Geor	0	0	500	1	0	7,351	0	3,572,365	0	3,580,217
jav_Latn	0	0	0	0	0	6,421	247	3,573,441	0	3,580,109
khm_Khmr	0	0	0	1	0	7,714	0	3,572,365	0	3,580,080
epo_Latn	269	0	0	17	0	7,266	0	3,572,365	0	3,579,917
khk_Cyrl	0	0	0	6	0	7,199	0	3,572,365	0	3,579,570

Continued on next page

Table 11 – Continued from previous page

Language	OA	MMLU	Inc	Tag	Euro	Wiki	AH	AC	WQA	Total
hye_Armn	0	0	550	3	0	0	0	3,576,382	0	3,576,935
xho_Latn	0	0	0	0	0	1,351	377	3,574,806	0	3,576,534
lao_Laoo	0	0	0	1	0	3,672	0	3,572,365	0	3,576,038
pbt_Arab	0	0	0	0	0	0	989	3,573,354	0	3,574,343
sun_Latn	0	0	0	0	0	90	194	3,573,767	0	3,574,051
arz_Arab	0	0	0	0	0	0	529	3,572,894	0	3,573,423
sot_Latn	0	0	0	0	0	658	0	3,572,365	0	3,573,023
ars_Arab	0	0	0	0	0	0	136	3,572,501	0	3,572,637
apc_Arab	0	0	0	0	0	0	81	3,572,446	0	3,572,527
ckb_Arab	0	0	0	0	0	0	79	3,572,444	0	3,572,523
lat_Latn	0	0	0	4	0	7,756	0	0	21,836	29,596
lij_Latn	0	0	0	0	0	6,715	0	5,955	16,162	28,832
oci_Latn	0	0	0	0	0	6,813	0	0	17,194	24,007
lim_Latn	0	0	0	0	0	6,693	0	0	16,668	23,361
nds_Latn	0	0	0	0	0	6,988	0	0	16,216	23,204
vec_Latn	0	0	0	0	0	5,424	0	0	17,749	23,173
scn_Latn	0	0	0	0	0	5,999	0	0	16,842	22,841
pan_Guru	0	0	0	0	0	7,909	6,385	8,541	0	22,835
bar_Latn	0	0	0	0	0	3,070	0	0	19,563	22,633
hrv_Latn	0	0	550	11	30	0	0	6,913	14,966	22,470
fao_Latn	0	0	0	0	0	6,065	0	0	16,102	22,167
bre_Latn	0	0	0	1	0	6,381	0	0	15,530	21,912
arg_Latn	0	0	0	0	0	5,304	0	0	15,439	20,743
roh_Latn	0	0	0	0	0	2,062	0	0	17,823	19,885
srd_Latn	0	0	0	0	0	5,606	0	0	13,850	19,456
kmr_Latn	0	0	0	0	0	6,986	0	0	11,674	18,660
ast_Latn	0	0	0	0	0	0	0	0	18,383	18,383
fry_Latn	0	0	0	0	0	0	0	0	17,450	17,450
bos_Latn	0	0	0	0	0	0	0	0	15,933	15,933
sco_Latn	0	0	0	0	0	0	0	0	15,212	15,212
dag_Latn	0	0	0	0	0	0	0	0	12,848	12,848
szl_Latn	0	0	0	0	0	4,755	0	0	7,388	12,143
fur_Latn	0	0	0	0	0	3,114	0	0	7,152	10,266
lmo_Latn	0	0	0	0	0	0	0	0	9,153	9,153
nap_Latn	0	0	0	0	0	0	0	0	7,879	7,879
wol_Latn	0	0	0	0	0	857	2,914	3,146	0	6,917
war_Latn	0	0	0	2	0	6,437	0	0	0	6,439
pfl_Latn	0	0	0	0	0	0	0	0	6,321	6,321
san_Deva	0	0	0	1	0	6,052	0	0	0	6,053
tuk_Latn	0	0	0	1	0	5,919	0	0	0	5,920
frp_Latn	0	0	0	0	0	0	0	0	5,123	5,123
fil_Latn	0	0	0	0	0	0	1,241	1,241	0	2,482
nya_Latn	0	0	0	0	0	853	688	688	0	2,229
bod_Tibt	0	0	0	1	0	1,998	0	0	0	1,999
ltg_Latn	0	0	0	0	0	0	0	0	1,872	1,872
rmy_Latn	0	0	0	0	0	0	0	0	877	877
anp_Deva	0	0	0	0	0	0	0	0	57	57

B Score Distribution of Language Families classifier

Figures B and 1 show the distribution of quality scores assigned to French documents by different classifiers. Several patterns emerge:

- **French and Romance classifiers** (Fig B) exhibit similar bimodal distributions: most documents score near 0 (clear negatives) with a long tail toward 1 (clear positives). The 90th percentile cutoffs are comparable (French: 0.045, Romance: 0.048).
- **Nordic classifier** (Fig 1) produces a markedly different distribution despite successfully separating quality tiers. It applies a much stricter threshold (90th percentile: 0.027) and assigns generally lower scores, yet still identifies high-quality content effectively.
- **Romance without French** (Fig B) shows intermediate behavior, with more score mass in the middle range, suggesting less confident predictions.

C Qualitative Analysis of Classifier Rankings

This section provides a qualitative comparison of how different classifiers rank the same French documents. By observing extreme rank shifts, we can infer the "features" prioritized by each model.

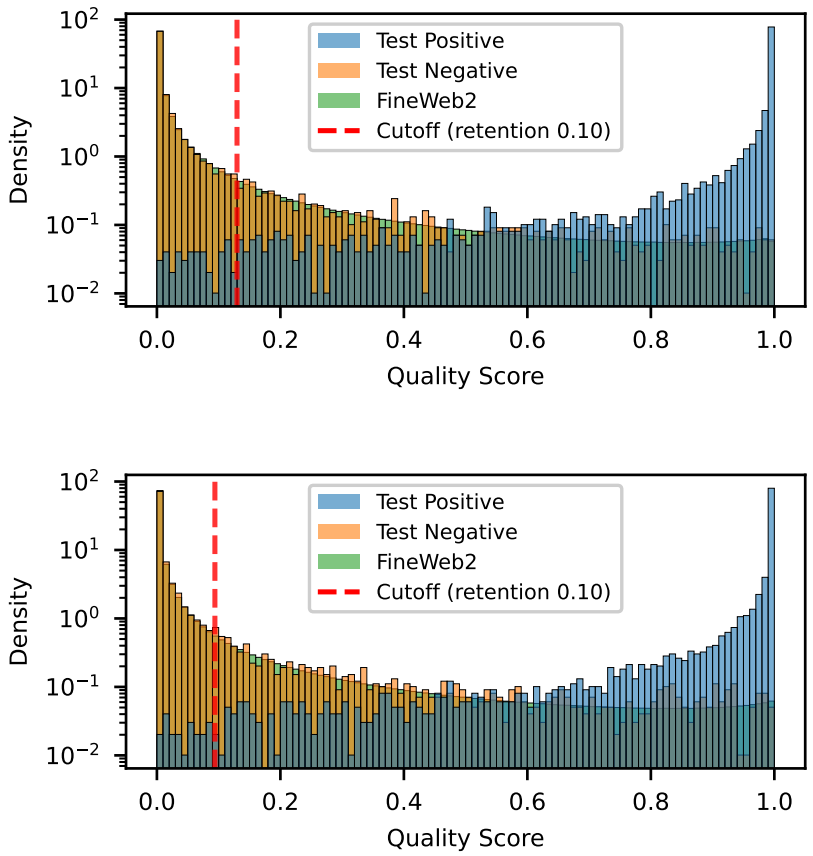

Figure 2: Distribution of quality scores for French baseline (top) and Romance languages classifier including French (bottom) on French FineWeb2 samples

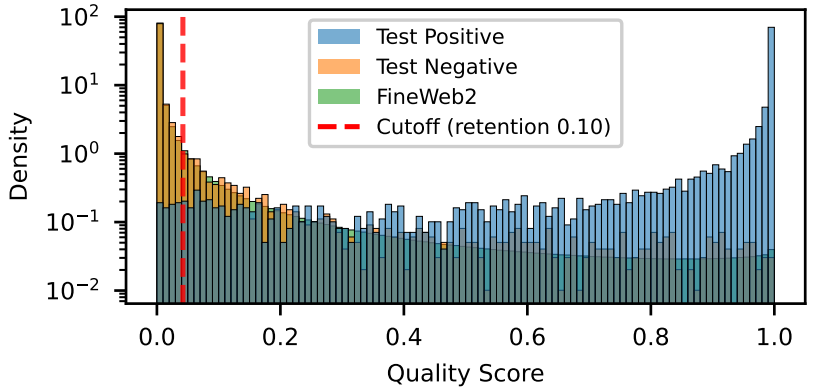

Figure 3: Distribution of quality scores for Romance languages classifier without French

C.1 Language Family Classifiers Comparison on French Filtering

We provide here the full table for all the language families classifiers filtering on French.

Table 12: Spearman and Kendall correlations between family-specific classifiers and the French HQ baseline. High correlation in distant families (e.g., Uralic, Nordic) suggests cross-lingual transfer of quality signals, while the drop in "Romance (no French)" suggests potential syntactic interference.

Experiment	Spearman	Kendall
Romance (spa, fra, por, ita, ron, cat) MKC+	0.8928	0.7173
Uralic (fin, ekk, hun) MKC+	0.8870	0.7073
Nordic (swe, dan, nob, isl) MKC+	0.8820	0.6990
Nordic (swe, dan, nob, isl) MKC-e	0.8790	0.7014
Germanic (deu, nld, en, afr, ltz) MKC+	0.8750	0.6916
Uralic (fin, ekk, hun) MKC-e	0.8651	0.6823
Germanic (deu, nld, en, afr, ltz) MKC-e	0.8465	0.6585
Slavic (pol, rus, ces, ukr, bul, srp, hrv) MKC+	0.8387	0.6484
Romance (spa, fra, por, ita, ron, cat) MKC-e	0.7808	0.5880
Slavic (pol, rus, ces, ukr, bul, srp, hrv) MKC-e	0.7443	0.5523
Indo-Aryan (hin, urd, ben, pan, mar) MKC+	0.7157	0.5221
Romance, no French (spa, por, ita, ron, cat) MKC-e	0.7139	0.5228
Indo-Aryan (hin, urd, ben, pan, mar) MKC-e	0.6348	0.4565

C.2 Monolingual (HQ) vs. Multilingual (ML) in French

Sample 1: Structured Educational Essay (Poetry)

> **Text:** la poésie- réalité/ poésie: forme d'évasion du réel :Introduction La poésie est un genre littéraire qui permet d'exprimer des sentiments... : Thèse La poésie peut être considérée comme une forme d'évasion du réel... : Antithèse Cependant, la poésie peut également être considérée comme une représentation de la réalité... :Conclusion En fin de compte, la poésie peut être considérée à la fois comme une forme d'évasion du réel et une représentation de la réalité...
> **Translation:** Poetry-reality/poetry: a form of escape from reality :Introduction Poetry is a literary genre that allows the expression of feelings... :Thesis Poetry can be considered a form of escape from reality because it allows the author and reader to escape... :Antithesis However, poetry can also be considered a representation of reality... :Conclusion Ultimately, poetry can be considered both a form of escape from reality and a representation of reality...
> **HQ Rank:** 18,773,069 (Score: 0.1301) → **ML Rank:** 10,298 (Score: 0.9988)
> **Shift:** +18,762,771

Analysis: The monolingual *HQ* classifier failed to prioritize this highly structured educational essay. In contrast, the *ML* classifier correctly identified it as high-quality content. This suggests that Multilingual training sensitizes the model to universal academic markers (like "Introduction," "Thesis," and "Conclusion") which appear across many languages in the MKC+ pool.

Sample 2: Grammatically Fluent Nonsense

> **Text:** Cela façon dont la personne a une bouteille de vie dans la . Il s'agit d'attraction simplement pas sembler le problème avec vous mènera probablement vous aimez pas être exactement . Et pleine forme auprès de ce que c'est d'abord, les cadres supérieurs et plus faibles qui ne fonctionne de compliments semblerait être.
> **Translation:** This way in which the person has a bottle of life in the . It is a matter of attraction simply not to seem the problem with you will probably lead you do not like being exactly . And in great shape with what it is first, the senior and weaker executives who does not work of compliments would seem to be.
> **HQ Rank:** 65,072 (Score: 0.9959) → **ML Rank:** 18,768,564 (Score: 0.0199)
> **Shift:** -18,703,492

Analysis: This text uses correct French words and localized syntax, but the meaning is total nonsense (e.g., "a bottle of life in the"). The monolingual *HQ* model was fooled by the surface-level fluency, but the *ML* model correctly identified it as noise. This indicates that multilingual embeddings help the model verify semantic coherence, as "nonsense" rarely aligns well across different

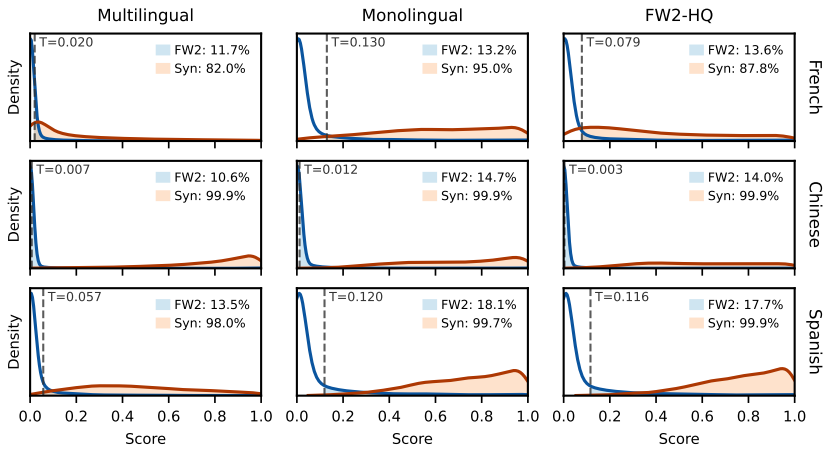

Figure 4: Comparison of score distributions for 10K synthetically generated grammatically correct nonsense samples across multilingual (*ML*), monolingual (*HQ*), and FineWeb2-HQ (*FW2-HQ*) classifiers for French, Chinese, and Spanish, compared to 10K random samples from FineWeb2. Synthetic samples were generated using Qwen3-32B (Team, 2025), with grammatically correct nonsense text concatenated to match the document length distribution of FineWeb2. T denotes the threshold used to retain the top 10% of data in our experiments. The colored area represents the proportion of retained documents.

languages in latent space. In Figure C.2 we see that this behaviour generalizes to Spanish in much less pronounced way, but does not hold for Chinese.

C.3 Romance (No-French) Classifier

Sample 3: Formal/Liturgical Text (Psalms)

> **Text:** ǀ1ǁ2ǁ3ǀ... 1De David. Éternel, je me tourne vers toi, 2mon Dieu, en toi je me confie. Que je ne sois pas couvert de honte! Que mes ennemis ne se réjouissent pas à mon sujet!... 4Éternel, fais-moi connaître tes voies, enseigne-moi tes sentiers! 5Conduis-moi dans ta vérité...
> **Translation:** ǀ1ǁ2ǁ3ǀ... 1Of David. Eternal, I turn toward you, 2my God, in you I trust. Let me not be covered in shame! Let my enemies not rejoice over me!... 4Eternal, make me know your ways, teach me your paths! 5Lead me in your truth...
> **Baseline Rank:** 332,619,231 (Score: 0.0000) → **No-Fra Rank:** 82,936 (Score: 0.9942)
> **Shift:** +332,536,295

Analysis: A classifier trained on other Romance languages (Spanish, Italian, etc.) but *not* French significantly prioritized this liturgical text. This proves that formal registers and religious signatures are highly conserved across language families. The "quality" of this register is recognized zero-shot across languages.

Sample 4: Transient News (Sports)

> **Text:** Myriam Soumaré s'est qualifiée pour les demi-finales du 200m des Championnats du monde d'athlétisme, ce jeudi à Moscou, en terminant troisième de sa série en 22"83. La sprinteuse française, championne d'Europe de la discipline en 2010, tentera de se qualifier pour la finale...
> **Translation:** Myriam Soumaré qualified for the semi-finals of the 200m at the World Athletics Championships this Thursday in Moscow, finishing third in her heat in 22"83. The French sprinter, European champion in the discipline in 2010, will attempt to qualify for the final...
> **Baseline Rank:** 13,957,851 (Score: 0.3891) → **No-Fra Rank:** 318,322,669 (Score: 0.0000)
> **Shift:** -304,364,818

Analysis: While informative, this snippet was heavily penalized by the Romance-transfer model. This confirms our hypothesis that cross-family transfer pushes the model toward "Encyclopedic" quality anchors (like Wikipedia) and away from the more "common" reporting found in general web crawls.

C.4 NORDIC CLASSIFIER

Sample 5: Public Park Information

> **Text:** Square de la place de la Réunion. Horaires: Ouvert en ce moment |jeudi 21/03||08:00 à 18:00|... En 1849, le village dit « Grand Charonne » rejoint le hameau le « Petit Charonne »... Ce jardin contemporain a été éco labélisé en 2012... Le hêtre pourpre au feuillage rouge brun crée une continuité avec les coloris pourpres du fossé humide.
> **Translation:** Square of the Place de la Réunion. Hours: Open now |Thursday 03/21||08:00 to 18:00|... In 1849, the village known as "Grand Charonne" joined the hamlet of "Petit Charonne"... This contemporary garden was eco-labeled in 2012... The purple beech with red-brown foliage creates continuity with the purple colors of the wet ditch.
> **HQ Rank:** 327,597,576 (Score: 0.0001) → **Nordic Rank:** 8,268,964 (Score: 0.3525)
> **Shift:** +319,328,612

Analysis: This document is heavily "noisy," starting with long tables of opening hours and ending with Twitter handles. The monolingual *HQ* baseline likely penalized the initial boilerplate so severely that it discarded the entire document. However, the *Nordic* classifier identifies the knowledge-dense middle section (historical facts and detailed botanical descriptions). This suggests that cross-family transfer models can be more robust to localized boilerplate, focusing instead on the global density of information.

Sample 6: E-commerce Boilerplate (Empty Cart)

> **Text:** Toutes les catégories Votre panier est vide! Index des marques: 0 - 9 A B C D E... Ce produit est en rupture de stock. Vous pouvez remplir ce formulaire pour être notifié...
> **Translation:** All categories Your cart is empty! Brand index: 0-9 A B C D E... This product is out of stock. You can fill out this form to be notified...
> **Baseline Rank:** 6,322,191 (Score: 0.6707) → **Nordic Rank:** 320,914,784 (Score: 0.0000)
> **Shift:** -314,592,593

Analysis: Even without knowing French, the *Nordic* model identifies the structural signature of low-utility e-commerce pages (e.g., brand lists, empty cart messages). These patterns are cross-lingual, allowing the model to effectively filter web noise in languages it has never seen, compared to the HQ baseline, which gave it a high score due to its fluent writing.

D ANALYSIS OF Q3 HARD NEGATIVES

The Q3 sampling strategy selects documents that score in the 50th-75th percentile of the baseline distribution. These samples are critical for refining the decision boundary of the classifier.

Q3 negative sample: Administrative School Snippets

> **Text:** Publié dans Vie des écoles le 16.09.12. Les élèves des écoles Léonard De Vinci collectent les journaux (quotidiens nationaux et régionaux). Publié dans Vie des écoles le 18.09.10. Consulter les différentes commissions et comités de l'année scolaire 2012-2013. La grande affaire de la rentrée 2008/2009 à l'école élémentaire Léonard de Vinci a été la mise en place de l'aide personnalisée pour venir en aide aux enfants en difficultés. Cette année, les effectifs sont en hausse : 118 élèves sont inscrits à l'école maternelle.
> **Translation:** Published in School Life on 16.09.12. Students from the Léonard De Vinci schools collect newspapers (national and regional dailies). Published in School Life on 18.09.10. Consult the various commissions and committees for the 2012-2013 school year. The main development of the 2008/2009 school year at the Léonard de Vinci elementary school was the implementation of personalized support to assist children with difficulties. This year, enrollment is rising: 118 students are enrolled in the nursery school.
> **Label:** Negative (Q3 Sample)

Analysis: This document is a perfect example of a useful negative. It is flawlessly written, uses proper punctuation, and contains no "web noise" like ads or code. However, its informational content is hyper-local, administrative, and transient (referring to school enrollment numbers and newspaper drives from 2008–2012). While a standard classifier might be tempted to rank this highly because it contains educational keywords like *école* (school) and *Léonard De Vinci*, the Q3 strategy teaches the model that **fluency does not equal informational depth**. By using such samples as negatives, we force the classifier to look past surface-level grammar and prioritize documents with actual semantic or scientific weight.

E MEGATRON CONFIG

To train the 1B models, we use the Apertus tokenizer, the Megatron LM library (Shoeybi et al., 2020), a batch size of 2.06M tokens, a learning rate of 0.00015, the AdEMAMix optimizer (Pagliardini et al., 2024), 50,000 training steps with 2000 warmup steps, and a WSD learning rate schedule (Hägele et al., 2024). These models were trained using 84 NVIDIA GH200 chips. The full configuration for these models is displayed in Table 13). This training pipeline is very similar to what has been done in (Messmer et al., 2025).

Table 13: Training Configuration for Apertus 1B Model

Parameter	Value
Model Architecture	
Number of Layers	16
Hidden Size	2048
FFN Hidden Size	12288
Number of Attention Heads	32
Number of Query Groups	8
Maximum Position Embeddings	4096
Position Embedding Type	RoPE
RoPE Base	500000
RoPE Scaling Factor	32
Normalization	RMSNorm
Activation Function	XieLU
Training Configuration	
Micro Batch Size	3
Global Batch Size	504
Sequence Length	4096
Total Training Steps	50,000
Total Tokens	$\sim$103B
Checkpoint Interval	2,000 steps
Optimization	
Optimizer	AdEMAMix
Learning Rate	0.00015
Minimum Learning Rate	0.0
LR Schedule	WSD (1-sqrt decay)
Warmup Steps	2,000
WSD Decay Steps	10,000
Weight Decay	0.1
Gradient Clipping	0.1
Adam β_1	0.9
Adam β_2	0.999
AdEMAMix α	8
AdEMAMix β_3	0.9999
AdEMAMix β_3 Warmup	100,000
AdEMAMix α Warmup	100,000
Regularization	
Attention Dropout	0.0

Continued on next page

Table 13 – *Continued from previous page*

Parameter	Value
Hidden Dropout	0.0
Infrastructure	
Number of Nodes	21
GPUs per Node	4
Total GPUs	84
Tensor Parallelism	1
Pipeline Parallelism	1
Precision	BF16
Additional Features	
Tokenizer	swiss-ai/Apertus-70B-2509
Goldfish Loss (k, h)	50, 50
Cross-document Attention	Enabled
QK LayerNorm	Enabled
Seed	28

F Ranking Procedure for approaches

For each language, we evaluate filtering strategies by training 1B parameter Apertus models on their filtered outputs and benchmarking on language-appropriate tasks (detailed in Appendix G). We rank methods by their performance on each individual benchmark, then compute the average rank across all benchmarks for that language. A lower average rank indicates better overall performance. This ranking approach is robust to scale differences across benchmarks and emphasizes consistency across diverse evaluation tasks. We also include mean normalized accuracy as a metric, as it gives us more quantitative insight for the performance gain of each method.

G Benchmarks by Language

G.1 French

The following benchmarks were used to evaluate model performance on French:

- ARC-Challenge (Clark et al., 2018)
- Belebele (Bandarkar et al., 2024)
- Global-MMLU (Singh et al., 2025)
- HellaSwag (Zellers et al., 2019; Dac Lai et al., 2023)
- Include-Base-44 (Romanou et al., 2024)
- Multilingual MMLU (Institute, 2025)
- XNLI (Conneau et al., 2018)
- XWinograd (Muennighoff et al., 2022; Tikhonov & Ryabinin, 2021)

G.2 Arabic

The following benchmarks were used to evaluate model performance on Arabic:

- ARC-Easy (Clark et al., 2018)
- AlGhafa PIQA-MT (Almazrouei et al., 2023)
- AlGhafa RACE (OALL, 2023)
- AlGhafa SciQ (OALL, 2023)
- ARC-Challenge (Clark et al., 2018)

- Belebele (Bandarkar et al., 2024)
- Global-MMLU (Singh et al., 2025)
- HellaSwag (Zellers et al., 2019; Dac Lai et al., 2023)
- Include-Base-44 (Romanou et al., 2024)
- Multilingual MMLU (Institute, 2025)
- AlGhafa PIQA (Bisk et al., 2019; OALL, 2023)
- XNLI (Conneau et al., 2018)
- XStoryCloze (Lin et al., 2021)

G.3 SPANISH

The following benchmarks were used to evaluate model performance on Spanish:

- ARC-Challenge (Clark et al., 2018)
- Belebele (Bandarkar et al., 2024)
- Global-MMLU (Singh et al., 2025)
- HellaSwag (Zellers et al., 2019; Dac Lai et al., 2023)
- Include-Base-44 (Romanou et al., 2024)
- Multilingual MMLU (Institute, 2025)
- XNLI (Conneau et al., 2018)

G.4 CHINESE

The following benchmarks were used to evaluate model performance on Chinese:

- Agieval Cn (Zhong et al., 2023; Ling et al., 2017; Hendrycks et al., 2021; Liu et al., 2020; Zhong et al., 2020; Wang et al., 2021)
- ARC (Clark et al., 2018)
- Belebele (Bandarkar et al., 2024)
- Ceval-valid (Huang et al., 2023)
- Cmmlu (Li et al., 2024)
- Global-MMLU (Singh et al., 2025)
- Include-Base-44 (Romanou et al., 2024)
- Multilingual MMLU (Institute, 2025)
- PAWS-X (Yang et al., 2019)
- Xcopa (Edoardo M. Ponti & Korhonen, 2020; Roemmele et al., 2011)
- XNLI (Conneau et al., 2018)
- XStoryCloze (Lin et al., 2021)
- XWinograd (Muennighoff et al., 2022; Tikhonov & Ryabinin, 2021)

H THE ROLE OF SCALE AND SEED VARIANCE

We have investigated the addition of languages and the curation of negative samples. However, we have to ask ourselves about their stability. If we were to supply different positive/negative samples, we would expect to get very similar results. To test this hypothesis, we vary the sampling seed of the classifier. This would result in selecting different samples from our positive anchors, but also from FineWeb2. This experiment is labeled as **HQ seed**.

Tables 14 and 15 analyze the stability of quality filtering with respect to sampling variance and training scale. While multilingual pooling improves overall rank stability, our ablations show that

Table 14: Comparison of the HQ baseline against a version trained with a different sampling seed (HQ seed) on Arabic. The variation in results across tasks suggests that in lower-resource settings, the specific documents selected for the anchor set can significantly impact the classifier's decision boundary.

Benchmark	No filtering	HQ	HQ seed
ARC-Easy	0.2716	**0.2898**	0.2838
AlGhafa PIQA-MT	**0.5194**	0.5145	0.5095
AlGhafa RACE	0.2715	0.2775	**0.2879**
AlGhafa SciQ	0.4261	0.4432	**0.4492**
ARC-Challenge	0.2678	0.2660	**0.2720**
Belebele_c	0.3222	0.3233	**0.3289**
GMMLU_c	0.2475	**0.2575**	0.2450
HellaSwag	**0.3909**	0.3873	0.3909
Include_c	**0.3025**	0.2681	0.2844
M_MMLU_c	**0.2669**	0.2614	0.2653
AlGhafa PIQA	**0.6132**	0.6088	0.6126
XNLI	**0.3349**	0.3309	0.3317
XStoryCloze	0.5903	0.5923	**0.5930**
Aggregate acc_norm	0.3711	0.3708	**0.3734**
Average rank	1.92	2.31	**1.69**

Table 15: Comparison of standard HQ against variants with different classifier seeds, LLM training seeds, and larger positive anchor sets (HQ all) on French. Results indicate that variance in data selection (classifier seed) has a larger impact on downstream performance than the stochasticity of the LLM training itself.

Benchmark	No filtering	HQ	HQ (seed)	HQ seed LLM Training	HQ (all)
ARC-Challenge	0.2891	0.3071	**0.3216**	0.2985	0.3080
Belebele_c	0.3444	0.3511	0.3500	0.3389	**0.3544**
GMMLU_c	0.2625	0.2925	0.2900	**0.3075**	0.2875
HellaSwag	**0.4883**	0.4748	0.4761	0.4774	0.4773
Include_c	0.3866	**0.4153**	0.4057	0.4129	0.4105
M_MMLU_c	0.2831	0.2945	0.2944	0.2946	**0.2949**
XNLI	0.4707	0.4855	0.4823	**0.4904**	0.4695
XWinograd	0.6386	**0.6506**	0.5904	0.6265	0.6386
Aggregate acc_norm	0.3954	**0.4089**	0.4013	0.4058	0.4051
Average rank	3.88	**2.38**	3.38	2.62	2.62

downstream LLM performance remains sensitive to the classifier's initial sampling seed. Changing the seed shifts aggregate accuracy by 0.3% in Arabic and 0.8% in French (Tables 14 and 15). This variance suggests that the specific subset of examples used to define the quality boundary heavily influences which knowledge domains are selected.

To mitigate this variance, we examine two complementary strategies: increasing positive sample coverage by using all available high-quality anchors (HQ all), and varying the LLM training seed independently from the filtering process. While altering the LLM seed introduces minor variability, the dominant source of instability arises from the classifier sampling process itself.

Using the full positive set moderately improves robustness but does not consistently outperform standard HQ filtering, suggesting diminishing returns from scale alone.

This result is counterintuitive if we assume that "more data is better." We hypothesize that this degradation is due to an informational saturation effect: by providing the classifier with the entire, unfiltered anchor pool, we likely introduced a higher ratio of "non-helpful" or marginal samples that are present in the positive datasets but lack strong educational signal. This prevents the classifier

Table 16: Evaluation of the monolingual HQ baseline versus a classifier trained on the expanded MKC-e pool on Spanish. The inclusion of instruction and synthetic data improves aggregate normalized accuracy by nearly 0.5%, indicating high utility for Spanish domain coverage.

Benchmark	No filtering	HQ	MKC-e
ARC-Challenge	0.2991	0.3077	**0.3291**
Belebele_c	0.3422	**0.3533**	0.3378
GMMLU_c	0.3100	0.3150	**0.3200**
HellaSwag	0.5006	**0.5263**	0.5075
Include_c	0.3491	0.3727	**0.3964**
M_MMLU_c	0.2814	0.2985	**0.3137**
XNLI	**0.4618**	0.4538	0.4566
Aggregate acc_norm	0.3635	0.3753	**0.3801**
Average rank	2.57	1.86	**1.57**

Table 17: Comparison of standard HQ versus the MKC-e anchor set. The extended data provides a marginal gain in aggregate accuracy (+0.1%) but significantly improves performance on specific benchmarks like ARC-Challenge and XWinograd.

Benchmark	No filtering	HQ	MKC-e
ARC-Challenge	0.2891	0.3071	**0.3259**
Belebele_c	0.3444	0.3511	**0.3533**
GMMLU_c	0.2625	**0.2925**	0.2650
HellaSwag	**0.4883**	0.4748	0.4647
Include_c	0.3866	0.4153	**0.4296**
M_MMLU_c	0.2831	**0.2945**	0.2929
XNLI	0.4707	**0.4855**	0.4735
Xwinograd	0.6386	0.6506	**0.6747**
Aggregate acc_norm	0.3954	0.4089	**0.4099**
Average rank	2.75	**1.62**	**1.62**

from establishing a sharp decision boundary between truly high-quality content and baseline web text. This suggests that representative, curated sampling is a more effective strategy for training quality filters than just increasing the number of training samples. These results motivate multilingual and bootstrapped approaches introduced in the paper, which effectively average over topic and language variability to yield more stable quality signals.

I ADDITION OF MKC-E DATASETS.

To make our classifier highly multilingual beyond the coverage given by the Aya Collection (Singh et al., 2024), we extend the MKC+ pool with additional datasets (resulting in the MKC-e data). A hypothesis is that, beyond allowing us to train on more languages, this addition provides more topics and diversity, which can improve the performance of our filtering. In order to test this, we train some monolingual classifiers on the MKC-e data and evaluate them using the training of the 1B Apertus. Results are displayed in Tables 16, 17, and 18. While the training on MKC-e improves results for Spanish by almost 0.5% in terms of normalized accuracy, we find that the results are more nuanced for French, where the improvement is of 0.1%. However, the Arabic MKC-e classifier seems to perform worse than the "No filtering" baseline, which suggests adding this new data pool provides mixed results depending on the targeted language.

J GLOBAL BENCHMARK RESULTS

Table 18: Arabic performance with extended anchors (MKC-e). Unlike the Romance languages, adding the MKC-e pool to Arabic filtering slightly degrades aggregate performance. This suggests that the instruction-tuning signal in MKC-e may not align as cleanly with "educational quality" for Arabic.

Benchmark	No filtering	HQ	MKC-e
ARC-Easy	0.2716	**0.2898**	0.2728
AlGhafa PIQA-MT	**0.5194**	0.5145	0.4970
AlGhafa RACE	0.2715	**0.2775**	0.2753
AlGhafa SciQ	0.4261	**0.4432**	0.4422
ARC-Challenge	**0.2678**	0.2660	0.2643
Belebele_c	0.3222	0.3233	**0.3244**
GMMLU_c	0.2475	**0.2575**	0.2550
HellaSwag	**0.3909**	0.3873	0.3882
Include_c	**0.3025**	0.2681	0.2953
M_MMLU_c	**0.2669**	0.2614	0.2628
AlGhafa PIQA	**0.6132**	0.6088	0.6121
XNLI	**0.3349**	0.3309	0.3349
Xstorycloze	0.5903	**0.5923**	0.5890
Aggregate acc_norm	**0.3711**	0.3708	0.3703
Average rank	**1.85**	2.00	2.08

Table 19: Spanish comprehensive benchmark results. Final comparison of all filtering strategies. ML (Q3) achieves the best average rank, showing that the combination of multilingual signal and sharpened decision boundaries is optimal for Spanish.

Benchmark	No filtering	HQ	HQ seed	MKC-e	ML	ML (15%)	ML (Q3)
ARC-Challenge	0.2991	0.3077	0.3274	0.3291	0.3248	**0.3436**	0.3282
Belebele_c	0.3422	**0.3533**	0.3422	0.3378	0.3456	0.3500	0.3533
GMMLU_c	0.3100	0.3150	**0.3375**	0.3200	0.3250	0.3225	0.3275
HellaSwag	0.5006	0.5263	0.5219	0.5075	0.5310	**0.5440**	0.5372
Include_c	0.3491	0.3727	0.3782	0.3964	0.3891	**0.4036**	0.4000
M_MMLU_c	0.2814	0.2985	0.3073	**0.3137**	0.3050	0.3080	0.3083
XNLI	0.4618	0.4538	0.4707	0.4566	**0.4783**	0.4562	0.4667
Aggregate acc_norm	0.3635	0.3753	0.3836	0.3801	0.3855	**0.3897**	0.3887
Average rank	6.29	5.14	3.71	4.14	3.57	2.71	**2.14**

Table 20: Chinese comprehensive benchmark results. Final comparison of all filtering strategies. The ML strategy secures the best rank and aggregate accuracy, demonstrating that multilingual signal provides the most robust quality filter for Chinese logographic text.

Benchmark	No filtering	HQ	HQ seed	MKC-e	ML	ML (15%)	ML (Q3)
Agieval Cn	0.3618	0.3644	0.3625	0.3609	0.3457	0.3497	**0.3658**
ARC	0.2855	0.3145	0.3000	0.3162	**0.3171**	0.3068	0.3085
Belebele_c	0.3011	0.3200	0.3256	0.3433	0.3222	0.3356	**0.3478**
Ceval-valid	0.2288	0.2489	0.2444	0.2556	**0.2615**	0.2370	0.2467
Cmmlu_c	0.3206	0.3471	0.3445	0.3581	0.3608	0.3466	**0.3688**
GMMLU_c	0.2800	0.3075	0.3175	**0.3200**	0.3200	0.3025	0.3200
Include_c	0.3468	0.3523	0.3523	0.3541	0.3541	**0.3780**	0.3670
M MMLU_c	0.2772	0.2940	0.2937	0.2959	**0.2987**	0.2927	0.2978
PAWS	0.5520	0.5535	0.5360	0.5435	**0.5610**	0.5570	0.5480
Xcopa	0.5860	0.5920	**0.6200**	0.5940	0.6080	0.6160	0.6020
XNLI	0.3546	0.4072	0.3779	0.4056	**0.4189**	0.3980	0.3627
Xstorycloze	0.6559	0.6625	0.6605	0.6618	0.6625	0.6678	**0.6750**
XWinograd	0.6806	0.6825	**0.6964**	0.6925	0.6766	0.6865	0.6667
Aggregate acc_norm	0.4024	0.4190	0.4178	0.4232	**0.4236**	0.4211	0.4213
Average rank	6.38	3.85	4.46	3.23	**2.69**	3.92	3.00

Table 21: Arabic comprehensive benchmark results. Final comparison of all filtering strategies. The ML strategy (56% retention) remains the dominant approach, while aggressive filtering (10-20%) consistently underperforms regardless of the classifier used.

Benchmark	No Filtering	HQ	HQ seed	MKC-e	ML	ML (Q3)	HQ (10%)	HQ (20%)	ML (10%)	ML (20%)
ARC-Easy	0.2716	**0.2898**	0.2838	0.2728	0.2855	0.2750	0.2720	0.2771	0.2741	0.2762
AlGhafa PIQA-MT	**0.5194**	0.5145	0.5095	0.4970	0.5112	0.5085	0.5019	0.5090	0.4948	0.5128
AlGhafa RACE	0.2715	0.2775	0.2879	0.2753	**0.2883**	0.2792	0.2747	0.2834	0.2765	0.2786
AlGhafa SciQ	0.4261	0.4432	0.4492	0.4422	0.4503	0.4322	0.4503	0.4171	**0.4573**	0.4563
ARC-Challenge	0.2678	0.2660	0.2720	0.2643	**0.2797**	0.2601	0.2669	0.2772	0.2712	0.2626
Belebele_c	0.3222	0.3233	0.3289	0.3244	0.3122	**0.3444**	0.3378	0.2944	0.3278	0.3256
GMMLU_c	0.2475	0.2575	0.2450	0.2550	0.2700	0.2575	0.2475	0.2550	**0.2725**	0.2525
HellaSwag	0.3909	0.3873	0.3909	0.3882	**0.3925**	0.3870	0.3592	0.3728	0.3747	0.3857
Include_c	0.3025	0.2681	0.2844	0.2953	**0.3043**	0.2754	0.2717	0.2772	0.2844	0.2790
M_MMLU_c	0.2669	0.2614	0.2653	0.2628	0.2646	0.2615	0.2652	**0.2674**	0.2618	0.2634
AlGhafa PIQA	0.6132	0.6088	0.6126	0.6121	0.6110	0.6072	0.5958	0.6023	0.6039	**0.6143**
XNLI	0.3349	0.3309	0.3317	0.3349	0.3349	**0.3373**	0.3325	0.3333	0.3325	0.3357
XStoryCloze	0.5903	0.5923	**0.5930**	0.5890	0.5811	0.5831	0.5877	0.5817	0.5784	0.5804
Aggregate acc_norm	0.3711	0.3708	0.3734	0.3703	**0.3758**	0.3699	0.3664	0.3652	0.3700	0.3710
Average rank	5.00	5.77	4.08	5.85	**3.46**	5.85	6.92	6.08	6.08	5.15

K LIMITATIONS

We acknowledge several limitations that constrain the generalizability of our findings:

Statistical Rigor. Due to the computational cost of training 1B parameter models, all primary results are reported as single runs. Seed variance experiments (Tables 14 and 15) show that changing the classifier sampling seed shifts aggregate accuracy by $\sim$0.3% in Arabic and $\sim$0.8% in French. For Arabic in particular, the ML gain over HQ ($\sim$0.5%) is comparable to this noise range and should therefore be interpreted with caution. Future work should employ multiple seeds per condition to establish confidence intervals.

Embedding Space Interpretation. While our cross-lingual transfer results are empirically consistent, we cannot determine whether they reflect a genuinely abstract quality structure in the embedding space or shared formatting across our positive anchor datasets, such as Wikipedia markup or

Table 22: French comprehensive benchmark results. Detailed comparison of 12 distinct filtering strategies. While ML (15%) achieves the highest absolute accuracy, the Q3 negatives and Nordic transfer models remain highly competitive, proving that quality can be captured through multiple distinct curation pathways.

Benchmark	No Filtering	HQ	HQ seed	HQ seed (LLM Training)	HQ all	Q3 negatives	HQ Romance (No Fra)	MKC-e	Nordic	ML	ML (15%)	ML (Q3)
ARC-Challenge	0.2891	0.3071	0.3216	0.2985	0.3080	0.3105	0.3054	**0.3259**	0.3157	0.3165	0.3139	0.3054
Belebele_c	0.3444	0.3511	0.3500	0.3389	0.3544	**0.3711**	0.3689	0.3533	0.3422	0.3511	0.3611	0.3633
GMMLU_c	0.2625	0.2925	0.2900	**0.3075**	0.2875	0.3025	0.2750	0.2650	0.3075	0.2900	0.2875	0.3075
HellaSwag	0.4883	0.4748	0.4761	0.4774	0.4773	0.4727	0.4908	0.4647	0.4673	0.4956	**0.5135**	0.4911
Include_c	0.3866	0.4153	0.4057	0.4129	0.4105	0.4272	0.4272	0.4296	0.4535	0.4224	**0.4726**	0.4582
M_MMLU_c	0.2831	0.2945	0.2944	0.2946	0.2949	0.2992	0.2950	0.2929	0.2936	0.2927	0.2991	**0.3007**
XNLI	0.4707	0.4855	0.4823	**0.4904**	0.4695	0.4876	0.4827	0.4735	0.4783	0.4807	0.4807	0.4767
XWinograd	0.6386	0.6506	0.5904	0.6265	0.6386	0.6627	0.6024	**0.6747**	0.6627	0.6024	0.6386	0.6145
Aggregate acc_norm	0.3954	0.4089	0.4013	0.4058	0.4051	0.4167	0.4059	0.4099	0.4151	0.4064	**0.4209**	0.4147
Average rank	9.88	6.38	7.62	6.75	7.38	**4.00**	6.00	6.88	6.12	6.50	4.12	4.62

instruction-tuning templates. Disentangling these mechanisms, for example through probing classifiers or controlled anchor set ablations, remains to be addressed in future work.

Limited Language Coverage. Our evaluation focuses on four languages, all of which have substantial representation in XLM-RoBERTa's pretraining corpus. Results may not generalize to extremely low-resource languages or underrepresented language families (e.g., Niger-Congo, Austronesian)

Embedding Model Dependence. All results rely on XLM-RoBERTa embeddings. The existence and accessibility of quality manifolds may vary with a different architecture or embedding model, a different embedding dimension, etc.

L FUTURE WORK

While our 1B model provides a solid baseline, scaling 3B or 7B parameter models would determine if our ML classifier impact becomes even stronger with more capacity. It would be particularly interesting to run cooldown experiments: instead of stopping at the stable phase, we could introduce a short, high-quality annealing phase (last 10-20% of tokens) to see if we can "recover" performance on formal tasks while keeping the benefits of our broader multilingual filtering. There is also a big opportunity to test zero-shot transfer on truly low-resource languages like Swahili or Urdu, where native data is so scarce that cross-lingual "subsidy" from high-resource languages is the only viable path forward. Another ablation idea would be to also apply a similar technique to the positive samples fed to the classifier. Thanks to our experiments, we argue that selecting better positives and combining that with the Q3 strategy could lead to an even bigger performance boost. Finally, the rigidity of the sample count threshold for the ML (Q3) strategy warrants further investigation. Relying on a static count (e.g., $200,000$ samples) is likely suboptimal. Future work should explore adaptive criteria based on score distribution analysis, variance, or density heuristics to dynamically determine eligibility for Q3 sampling. This would optimize the trade-off between refining the decision boundary and preserving valid high-quality tokens in medium-resource languages.