arXiv:2507.01785v1 [cs.CL] 2 Jul 2025

MuRating: A High Quality Data Selecting Approach to Multilingual Large Language Model Pretraining

Zhixun Chen[1]* **Ping Guo**[2] **Wenhan Han**[2] **Yifan Zhang**[2] **Binbin Liu**[2]
Haobin Lin[2] **Fengze Liu**[2] **Yan Zhao**[2] **Bingni Zhang**[2] **Taifeng Wang**[2]
Yin Zheng[2]† **Meng Fang**[3]†
[1]University of Technology Sydney [2]ByteDance
[3]University of Liverpool

Abstract

Data quality is a critical driver of large language model performance, yet existing model-based selection methods focus almost exclusively on English. We introduce MuRating, a scalable framework that transfers high-quality English data-quality signals into a single rater for 17 target languages. MuRating aggregates multiple English "raters" via pairwise comparisons to learn unified document-quality scores, then projects these judgments through translation to train a multilingual evaluator on monolingual, cross-lingual, and parallel text pairs. Applied to web data, MuRating selects balanced subsets of English and multilingual content to pretrain a 1.2 B-parameter LLaMA model. Compared to strong baselines, including QuRater, AskLLM, DCLM and so on, our approach boosts average accuracy on both English benchmarks and multilingual evaluations, with especially large gains on knowledge-intensive tasks. We further analyze translation fidelity, selection biases, and underrepresentation of narrative material, outlining directions for future work.

1 Introduction

Large Language Models (LLMs) have achieved remarkable performance across a wide range of tasks, and recent studies have consistently emphasized the critical role of high-quality pretraining data in driving these advances [5, 42, 7]. To improve data quality, various strategies have been adopted, such as deduplication [26, 1], heuristic and rule-based filtering [42, 37], and domain-aware sampling [56, 45]. While effective, these methods often rely heavily on manual heuristics and domain expertise, lacking a unified or principled framework for evaluating and selecting pretraining data. Moreover, they are typically applied as pre-defined or post-hoc filters, limiting their adaptability to downstream performance. In response, model-based data selection approaches have emerged, aiming to learn data quality judgments from examples or auxiliary supervision. These methods utilize different model architectures and data selection criteria. For instance, DCLM [28] trains a FastText classifier [22] using high-quality samples from OH2.5 and Reddit ELI5 as positive supervision, while treating Common Crawl web data as negatives. Other approaches such as AskLLM, QuRater, and the FineWeb-Edu classifier [43, 55, 30] employ prompt-based evaluation criteria using various LLMs to assess the quality of input samples.

Data selection beyond English remains a profound challenge [13, 25]. While model-based data selection methods have demonstrated effectiveness in improving training quality, they have been developed almost entirely for English, leaving a critical gap in multilingual data quality assessment. As LLMs are increasingly applied in diverse linguistic contexts, there is a growing need for selection

[*] Work done during Zhixun's internship at ByteDance. †Corresponding Authors.

strategies that extend beyond English. A recent attempt Fineweb2-hq [31] train language-specific raters using benchmark datasets as positive supervision and general pretraining corpora as negatives, following a strategy similar to DCLM. However, this approach trains separate raters per language, demanding large, human-curated positive examples that simply do not exist for many low-resource languages. Additionally, using benchmark-derived data poses a risk of contamination.

In this work, we introduce MuRating, a two-stage, translation-and-pairwise framework for multilingual data-quality estimation. We begin by aggregating multiple state-of-the-art English raters via majority-vote pairwise comparisons, fitting a Bradley–Terry model [4] to learn a single, unified quality scorer. Next, we translate scored English document pairs into each of 17 target languages and construct monolingual, cross-lingual, and parallel pairs—projecting original preference labels onto translated comparisons and assigning neutral labels to parallel translations. This design yields one multilingual evaluator that preserves English-derived quality signals while remaining language-agnostic.

We apply MuRating framework to fine-tune a MuRater model to annotate 1.5 trillion English and 3 trillion multilingual web tokens, selecting 200 billion English and 300 billion multilingual tokens for pretraining a 1.2 B-parameter LLaMA model. Compared to strong baselines—uniform sampling (+50 % data), QuRater, AskLLM, DCLM—our selection yields an average gain of 1 to 3.4 points on twelve English benchmarks (including ARC, HellaSwag, MMLU) and 1.8 points on a diverse multilingual suite (XCOPA, XNLI, Flores, plus native MMLU-variants). We further assess translation fidelity via human evaluation, examine the impact of cross-lingual and parallel data, and compare different score transfer approaches.

Our contributions are as follows:

- Unified English rater aggregation. We consolidate four distinct English quality raters via a Bradley–Terry pairwise framework, producing a single, robust scoring model.
- Translation-based multilingual transfer. We show how to project English pairwise judgments into monolingual, cross-lingual, and parallel pairs across 17 languages, enabling language-agnostic quality evaluation.
- Scalable pretraining gains. MuRating's selected 500 billion-token corpus yields substantial improvements over state-of-the-art baselines on both English and multilingual LLM benchmarks.
- Analysis and open resources. We assess translation fidelity, address selection biases, and will release our prompts, code, and data to facilitate further research.

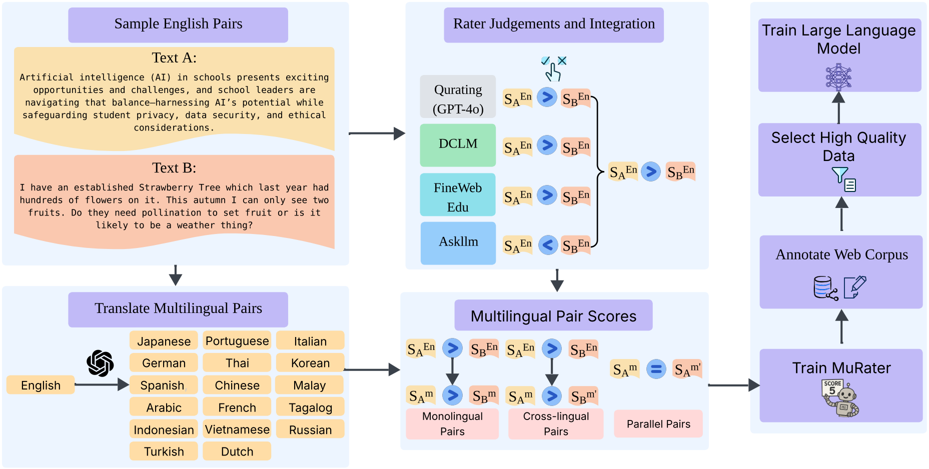
2 Related Work

Data Selection. Data selection is essential in constructing high-quality pretraining corpora for LLMs and typically falls into three main categories: deduplication, heuristic-based filtering, and LLM-guided quality evaluation. Deduplication, applied early, removes exact or near-duplicate documents to reduce redundancy and improve generalization [26]. More advanced fuzzy and semantic methods filter syntactically or semantically similar content [20, 1], which is crucial at scale to avoid training instability and performance degradation [59, 42].

Heuristic filtering uses rules or lightweight models to exclude low-quality text, such as short, repetitive, or toxic content [25, 53, 38, 47]. While handcrafted heuristics can be effective, they often have limited generalization and inefficiency, prompting the use of simple classifiers, perplexity scores, or importance sampling [7, 50, 57, 34]. However, these approaches may unintentionally favor simplistic or repetitive content, which can diminish the diversity and informativeness of the dataset.

In contrast, LLM-guided quality scoring directly leverages language models to evaluate data along dimensions like factuality and coherence [17, 43]. Frameworks such as QuRating and FineWeb-Edu prioritize educational content using multi-criteria assessment [55, 30], while Dataman [39] and FIRE [58] extend this to domain-aware or reliability-sensitive filtering. Despite their advancements, recent approaches depend heavily on GPT-style judgments, potentially introducing model-specific biases.

Multilingual Pretraining. Efforts to construct multilingual datasets for multilingual LLM pretraining have followed similar strategies to those used for English, incorporating deduplication and heuristic-based filtering techniques. Prominent corpora such as mC4 [60], RedPajama [53], CulturalX [35],



Text A: Artificial intelligence (AI) in schools presents exciting opportunities and challenges, and school leaders are navigating that balance—harnessing AI's potential while safeguarding student privacy, data security, and ethical considerations.

Text B: I have an established Strawberry Tree which last year had hundreds of flowers on it. This autumn I can only see two fruits. Do they need pollination to set fruit or is it likely to be a weather thing?

Figure 1: Overview of the MuRating pipeline: English document pairs are first annotated using various data selection methods and unified, then translated into multiple languages to create diverse multilingual pairs. These are used to train the MuRater model, which scores large-scale web data. The top 10% of scored data is selected to train an LLM, yielding superior performance compared to state-of-the-art sampling baselines.

HPTL [12], and FineWeb-2 [37] leverage these methods to scale multilingual resources, ranging from a few dozen to thousands of languages, significantly enhancing cross-lingual performance in multilingual LLM pretraining. To mitigate the issue of data scarcity for low-resource languages, TransWeb-Edu [52] addresses this by translating high-quality English data into multiple languages. However, there has been limited research on model-based data selection in the context of multilingual LLM pretraining. Recently, an initial approach [31] introduced a model-based selection method to refine the FineWeb-2 dataset by training language-specific classifiers, using multilingual benchmark datasets as positive examples and web corpus data as negative examples. However, this method relies on the availability of high-quality samples from existing multilingual benchmarks, which presents scalability challenges and limits its applicability to low-resource or underrepresented languages. Additionally, this approach may risk contaminating downstream evaluation tasks with biased data.

3 Methodology

Our approach consists of two stages: (1) consolidating multiple English-language quality raters into a single, unified scorer via pairwise comparisons, and (2) transferring that scorer to a multilingual setting through translation-based alignment and cross-lingual regularization. We introduce a two-step method: first, we integrate existing English corpus quality raters; second, we transfer their rating capability to a multilingual setting.

3.1 Integration of English Raters

To consolidate quality judgments from multiple pre-existing English raters, we employ a pairwise comparison framework grounded in statistical preference modeling. Let (t_A, t_B) denotes a pair of texts randomly sampled from a large corpus, and let N be the set of raters. Each rater $n \in N$ assigns a scalar score to both texts, denoted as S_A^n and S_B^n, reflecting the rater's estimation of the quality of t_A and t_B, respectively.

We define a binary preference for each rater: if $S_A^n > S_B^n$, we consider that rater n prefers text t_A over t_B, and vice versa. Based on the collective preferences of all raters, we compute an empirical confidence score $P_{A>B}$ indicating how likely t_A is preferred over t_B:

$$P_{A>B} = \frac{1}{|N|} \sum_{n \in N} \mathbb{I}[S_A^n > S_B^n], \quad P_{A>B} \in [0, 1], \tag{1}$$

where $\mathbb{I}[\cdot]$ is the indicator function that equals 1 when the condition is true and 0 otherwise. This score quantifies the relative preference strength of t_A over t_B across all raters.

To construct a large-scale preference dataset, we apply this scoring procedure across a wide set of sampled text pairs. This process yields a judgment dataset: $\mathcal{J} = \{(t_A, t_B, P_{A>B})\}$ consisting of text pairs and the estimated probability of preference.

To convert these pairwise comparisons into continuous scalar quality scores, we employ a learning framework based on the Bradley-Terry model [4]. Let $s_\theta(t)$ denote the learnable scalar quality score of text t, parameterized by θ. We adopt a binary cross-entropy loss function, following the formulation proposed in [55], which is analogous to the reward model training paradigm in Reinforcement Learning from Human Feedback (RLHF) [36], but without incorporating user prompts or conditioning on input queries:

$$\mathcal{L}_\theta = \mathop{\mathbb{E}}_{(t_A, t_B, p_{B \succ A}) \in \mathcal{J}} \left[-p_{B \succ A} \log \sigma(s_\theta(t_B) - s_\theta(t_A)) - (1 - p_{B \succ A}) \log \sigma(s_\theta(t_A) - s_\theta(t_B)) \right], \quad (2)$$

where, $\sigma(\cdot)$ denotes the sigmoid function, and $p_{B \succ A} = 1 - P_{A>B}$ is the empirical probability that t_B is preferred over t_A. This formulation encourages the model to assign higher scores to texts that are consistently preferred in the pairwise judgments.

After training, the model outputs a scalar quality rating for each document. These scores are treated as logits over the dataset and are used for quality-based sampling, where a subset of high-quality texts is selected based on their relative scores.

3.2 Multilingual Data Quality Rater

3.2.1 Translation-Based Alignment of Multilingual Preferences

To extend data quality scoring from English to a set of target languages M, we adopt a translation-based strategy. Building on the scored English text pairs introduced in the previous section, we translate each document pair (t_A^{en}, t_B^{en}) into a target language $m \in M$. For each pair, we compute a confidence score $P_{A^{en} > B^{en}}$ following Equation 1, and then directly transfer this preference to the translated pair by assuming $P_{A^m > B^m} \approx P_{A^{en} > B^{en}}$.

This assumption is based on the premise that translation preserves both the semantic content and the relative quality between text pairs. Prior work QuRating [55] highlights that pairwise comparisons offer increased stability when evaluating text quality. In multilingual settings, pointwise scoring—where absolute quality scores are assigned to individual texts—tends to be more susceptible to subtle changes in tone or phrasing introduced during translation, which can compromise the consistency of the supervision signal. In contrast, pairwise supervision is inherently more robust to such translation-induced variations. As long as the relative ranking between the texts remains consistent (i.e., t_A^{en} continues to be preferred over t_B^{en} after translation), the corresponding translated pair (t_A^m, t_B^m) remains a valid training example. This robustness makes pairwise comparisons a more reliable and effective framework for training quality evaluation models in multilingual contexts.

3.3 Cross-Lingual and Language-Agnostic Alignment

While the previous section addressed only in-language supervision—i.e., training on text pairs (t_A^m, t_B^m) where both documents are in the same language m—this setup alone is insufficient to guarantee language-agnostic scoring behavior. To promote consistency in quality assessments across languages, we augment our training dataset with both cross-lingual and parallel text pairs.

For cross-lingual supervision, we generate mixed-language pairs by translating t_A and t_B into different target languages, resulting in pairs of the form $(t_A^m, t_B^{m'})$ with $m \neq m'$. The original English pairwise preference score is then transferred to these cross-lingual pairs by assuming $P_{A^m > B^{m'}} \approx P_{A^{en} > B^{en}}$.

In addition, we introduce parallel pairs to explicitly regularize the model's behavior on semantically equivalent content across languages. Given a text t_A^m and its direct translation $t_A^{m'}$ into another language m', we form the pair $(t_A^m, t_A^{m'})$ and assign a neutral preference score, i.e., $P_{A^m > A^{m'}} \approx 0.5$. This reflects the expectation that both texts, despite being in different languages, convey identical semantic meaning and should be treated as equally quality.

Formally, these neutral-pair constraints act as a regularization signal that aligns the model's internal representation of quality across languages:

$$\mathcal{L}_{\text{parallel}} = \mathop{\mathbb{E}}_{(t_A, t_{A'}) \in \mathcal{J}'} \left[-\log \sigma \left(s_\theta(t_A^m) - s_\theta(t_A^{m'}) \right) - \log \sigma \left(s_\theta(t_A^{m'}) - s_\theta(t_A^m) \right) \right], \tag{3}$$

where $J^{'}$ is the datasets of parallel pairs. This formulation encourages the model to minimize score divergence between translations while still preserving the ability to differentiate documents of genuinely different quality in the broader training set.

3.3.1 Multilingual Rater Objective

The final loss function is a combination of the original pairwise loss from same-language and cross-language comparisons, along with the parallel text regularization term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pairwise}} + \lambda \cdot \mathcal{L}_{\text{parallel}}, \tag{4}$$

where λ is a tunable hyperparameter balancing cross-lingual consistency and discrimination. This joint training approach allows us to construct a multilingual quality rater that is robust, consistent across languages, and sensitive to relative quality differences.

3.3.2 Training the Rater Model

To build a high-quality multilingual rater, we sample 300,000 English text pairs and annotate them using four rating methods. For GPT-4o-based annotation, we prompt the model in both directions—(t_A, t_B) and (t_B, t_A)—multiple times to mitigate order bias. The final confidence score $P_{A>B}$ is computed by averaging predicted preference probabilities. For other raters (AskLLM, FineWeb-Edu-Classifier, DCLM), we collect individual scores and then integrate pairwise preferences using Equation 1.

We translate English pairs into 150,000 monolingual and 150,000 cross-lingual pairs, where languages portion is balanced across languages. Additionally, 75,000 texts are translated into two different languages to form parallel pairs $(t_A^m, t_A^{m'})$, which are assigned a neutral score of $P_{A>A'} = 0.5$ to promote consistency across languages for semantically equivalent content.

The final MuRater training set includes 75,000 English, 150,000 monolingual, 150,000 cross-lingual, and 75,000 parallel pairs. We adopt QuRater's training setup [55], applying a confidence margin to all but the parallel examples.

We fine-tune an encoder-based model following the BGE-M3 architecture [6], adding a linear head to predict quality ratings. The resulting rater achieves over 93% accuracy on held-out and 97% on held-in pairwise judgments, demonstrating strong multilingual preference modeling. Implementation details, including tokenizer settings and hyperparameters, are provided in Appendix A.

4 Experiments

4.1 Experimental Setups

Dataset construction. We build on the deduplication and heuristic-filtering pipelines of SlimPajama [46] and FineWeb [30] to assemble a large web-crawl corpus. It comprises 1.5 trillion English tokens plus 3 trillion tokens across 17 additional languages (Arabic, Chinese, Dutch, French, German, Indonesian, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Thai, Turkish, Vietnamese, Malay, Tagalog), all drawn from Fineweb-2 [37]. We then apply our multilingual rater to assign pairwise-derived quality scores to every document. Although scoring trillions of tokens is compute-intensive, it parallelizes efficiently across GPUs, and batching strategies reduce overhead in practice. Corpus statistics are detailed in Appendix B.1.

Baselines. For English-only experiments, we consider several established data quality raters: **QuRater**, which selects data based on educational value [55]; **AskLLM**, which follows the prompt design in [43] using Flan-T5-XXL [8]; the **FineWeb-Edu** Classifier[2], trained on 450K LLaMA3-70B-Instruct [3] labels to identify educational content; and **DCLM**[4], a fastText classifier trained on high quality dataset to differentiate between informative and low-quality web content.

For the multilingual setting, we adapt the QuRater framework to create **QuRater-M**, using GPT-4o to annotate multilingual pairs via relative preference, following the same training pipeline as its English counterpart. Additionally, we introduce **MuRater(M)** as a baseline, which involves sampling multilingual pair data, translating it into English for rating, and projecting the scores back to the original languages. Parallel and cross-lingual pairs are also incorporated to enhance the rater training.

Finally, we include a **Uniform** baseline for both settings, which randomly samples data (50% more than others). Details are provided in Appendix B.2.

Training Setup. We train a randomly initialized language model based on the LLaMA architecture [16] for a single epoch over the training corpus, with data presented in a randomly shuffled order. The model comprises 1.2 billion parameters and employs a standard transformer architecture [51] augmented with rotary position embeddings (RoPE) [48]. To accommodate the multilingual setting, we extend the tokenizer vocabulary through retraining on the multilingual corpus. Comprehensive architectural and tokenizer details are provided in Appendix B.3.

For English-language experiments, we select top scored 200 billion tokens from the full pool of 1.5 trillion tokens across all baseline methods and MuRater. In the multilingual setting, we apply our multilingual rater and the QuRater-M baseline to select the top 10% of tokens per language, resulting in approximately 300 billion tokens in total. These selected multilingual tokens are then combined with the 200 billion English tokens to yield a 500-billion-token training corpus.

Evaluation Benchmarks. We assess the performance of our pretrained models using the `lm-evaluation-harness` framework [14]. For the English-only evaluation, we consider a suite of ten tasks spanning multiple linguistic competencies. These include six reading comprehension benchmarks—ARC-Easy, ARC-Challenge [10], SciQ [54], LogiQA [29], TriviaAQ[21] and BoolQ [9]; four commonsense reasoning tasks—HellaSwag [61], PIQA [3], OpenBookQA [32] and WinoGrande [44]; and two knowledge-intensive tasks—Natural Questions (NQ) [24] and MMLU [19]. For MMLU, we follow [2] and employ the `lighteval` variant to ensure more consistent and reliable comparisons.

For the multilingual evaluation, we utilize translated versions of several English benchmarks in addition to multilingual-native datasets [18]. The evaluation suite includes translated ARC-Easy, ARC-Challenge, and HellaSwag, MMLU, StoryCloze [33] along with XCOPA [40], XNLI [11], XWinograd [49], BMLAMA [41] and FLORES [15], We further incorporate native language benchmarks—CMMLU [27], VMLU [5], IndoMMLU [23], JMMLU[6], and AMMLU[7]—to construct a localized multilingual variant of the MMLU test set, denoted as MMLU_L. These datasets collectively evaluate the model's capabilities in cross-lingual comprehension, reasoning, and translation resilience. A detailed overview of the benchmarks and their language coverage is provided in Appendix B.4.

4.2 Main Results

4.2.1 Multilingual Results

The results in Table 1 and Figure 2 demonstrate that MuRater consistently outperforms the multilingual baseline QuRater-M across most benchmarks by 1.8 percent, particularly excelling on knowledge-intensive and reading comprehension tasks such as ARC and MMLU. This indicates that MuRater is more effective in selecting high-quality multilingual pretraining data. Among its two variants, MuRater(E), which translates rated English data into multilingual counterparts, achieves

[2] `https://huggingface.co/HuggingFaceFW/fineweb-edu-classifier`

[3] `https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct`

[4] `https://huggingface.co/mlfoundations/fasttext-oh-eli5`

[5] `https://vmlu.ai/`

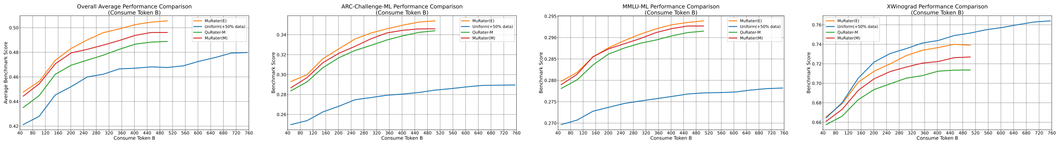
[6] `https://huggingface.co/datasets/nlp-waseda/JMMLU`

[7] `https://huggingface.co/datasets/Hennara/ammlu`

Table 1: Performance of different data selection strategies across downstream tasks under mixing 200B English and 300B multilingual tokens. **MuRater(M)** denotes training with multilingual pairs translated into English for scoring, while **MuRater(E)** uses rated English data translated into multilingual pair form. Best results within each setting are shown in **bold**.

Selection Method	MMLU_L	ARC_C_ML	ARC_E_ML	Flores	Hellaswag_ML	MMLU_T	XCOPA	XNLI	StoryCloze_ML	XWinograd	BMLAMA	Average
Uniform(+*50% data*)	29.98	28.74	49.03	46.66	44.56	27.83	64.60	42.33	**69.28**	**76.40**	48.55	48.00
QuRater-M	30.86	33.89	57.53	46.60	46.77	29.18	62.97	44.39	65.86	71.21	45.84	48.65
MuRater(M)	31.86	34.26	58.21	**47.43**	46.54	29.28	64.43	42.07	67.08	72.92	50.13	49.47
MuRater(E)	**31.96**	**35.01**	**58.98**	47.27	**47.11**	**29.40**	**65.73**	**44.46**	67.67	74.13	**52.97**	**50.43**

better performance than MuRater(M), suggesting that transferring rating capabilities from English to other languages via translation is a more efficient strategy. Despite its overall advantage, MuRater shows relatively weaker results on narrative and context-rich tasks like StoryCloze-ML and XWinograd, where the Uniform baseline performs better. These tasks require nuanced contextual comprehension and familiarity with diverse narrative styles, which are better captured by larger and more varied training data. MuRater's preference for fact-based, educationally relevant content boosts performance in QA-style evaluations but may lead to underrepresentation of narrative-rich samples. Furthermore, the Uniform baseline benefits from 50% more data, which contributes to its relative strength on narrative-oriented tasks. Detailed language-specified performance breakdowns are provided in Appendix D.



(a) Average performance (b) ARC-Challenge-ML (c) MMLU-ML (d) XWinograd

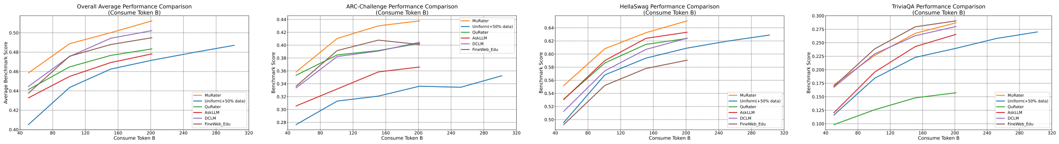
Figure 2: Performance of different selection methods on ARC-Challenge-ML, MMLU-ML, XWinograd, and the overall average across all tasks during training on 200B English + 300B multilingual tokens.

4.2.2 English-only Results

Table 2: Performance of different selection method over all different downstream tasks. Best results of each task category is marked in black. Detailed results are performed in Appendix D.

Selection Method	Reading Comprehension *(6 tasks)*	Commonsense Reasoning *(4 tasks)*	World Knowledge *(2 tasks)*	Average *(12 tasks)*
Uniform (+*50% data*)	43.93	59.06	20.36	48.70
Askllm	42.83	58.40	20.21	47.82
DCLM	46.00	58.99	22.37	50.23
FineWeb_Edu	45.71	57.49	22.00	49.49
QuRater	43.54	58.58	20.47	48.33
MuRater	**47.13**	**59.95**	**22.53**	**51.23**

The results in Table 2 indicate that our proposed rater successfully consolidates the strengths of existing rating methodologies, leading to consistent improvements in pretrained model performance across all categories of evaluation tasks. Baseline comparisons reveal that each selection method exhibits distinct preferences for data, which translate into varying levels of effectiveness on different downstream tasks. For instance, as shown in Figure 3, DCLM yields strong results on HellaSwag but underperforms on ARC-Challenge. Conversely, QuRater achieves competitive performance on ARC-Challenge but demonstrates poor results on TriviaQA. In contrast, our rater integrates the advantages of these methods and achieves robust performance across nearly all benchmarks, outperforming other data selection baselines by margins ranging from 1 to 3.4 percent. The model trained with our rater consistently achieves superior results on all tasks throughout the training process. This indicates more stable and efficient learning, further validating the effectiveness of our data selection approach in enhancing the quality of LLM pretraining.



(a) Average performance (b) ARC-Challenge (c) HellaSwag (d) TriviaQA

Figure 3: Performance of different selection methods on ARC-Challenge, HellaSwag, TriviaQA, and the overall average across 12 tasks during training on 200B English tokens

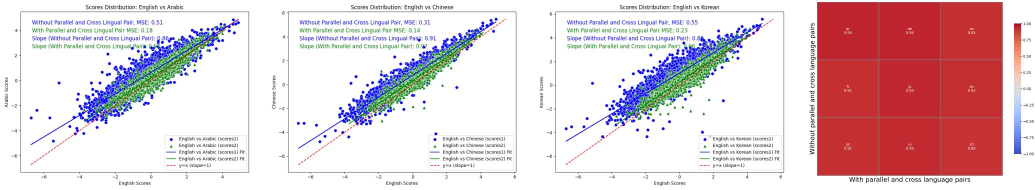
4.3 Ablation Study

4.3.1 Effectiveness of Cross-Lingual and Parallel Pair Integration

Incorporating cross-lingual pairs and parallel translations during training significantly improves the consistency of quality scoring across languages. To validate this, we assess multilingual raters on parallel corpora—semantically equivalent texts in different languages. In an ideal scenario, a language-agnostic rater would assign same scores for parallel language texts, yielding a slope of one and minimal mean squared error (MSE) between the score sequences of the corresponding texts. As shown in Figure 4, our alignment-based training yields models with lower MSE and slopes closer to one, demonstrating improved cross-lingual consistency.

These findings highlight the importance of modeling interlingual relationships. By leveraging cross-language comparisons and parallel data, the rater learns language-invariant quality standards, enabling more reliable multilingual evaluation. A qualitative case study in Appendix E further supports this, showing that high-rated texts consistently exhibit greater fluency, coherence, and instructional value across languages. We further examine how the selected data is distributed across semantic domains in different languages. Detailed results are provided in Appendix B.1.

In addition, we sample 10,000 annotated samples of nine languages using both aligned and unaligned raters. We compute the Kendall Tau coefficients between the resulting score sequences and visualize them in Figure 4d. The strong correlation indicates that alignment mainly standardizes absolute score values across languages without affecting the relative ranking. Moreover, selecting the top 10% high-quality samples from both configurations yields a 92% overlap, confirming that alignment preserves core evaluative behavior while enabling more consistent cross-lingual interpretation.



(a) Arabic (b) Chinese (c) Korean (d) Kendall Tau Corr.

Figure 4: Scatter plots of scores assigned by multilingual raters to 10,000 parallel documents across various languages. Green points represent ratings from raters trained with alignment using parallel and cross-lingual pairs, while blue points indicate scores from unaligned raters. The Kendall Tau matrix illustrates the rank correlation between the two raters' scoring outputs of 10,000 documents of different languages.

4.3.2 Translation Quality Asseement

We assess the quality of our translations through human evaluation. Expert annotators are provided with 50 pairs of source and translated texts and asked to rate translation quality on a scale from 1 to 5: scores 1 indicate severely flawed translations, 2 corresponds to incorrect translation with notable errors, 3 denotes generally correct translations with some errors, 4 represents good translations with

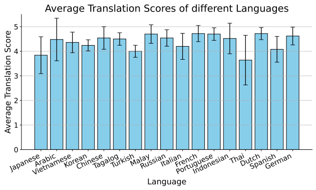


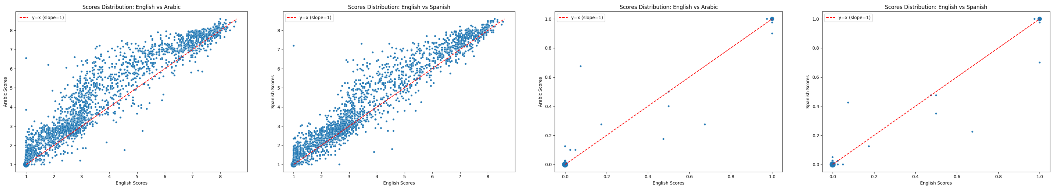
Figure 5: Translation average scores of various languages.

minor issues, and 5 corresponds to near-perfect translations. The evaluation criteria are detailed in Appendix C.1. As shown in Figure 5, the overall translation quality of GPT-4o is high, with most languages achieving average scores above 4. Notably, performance on Japanese and Thai is comparatively lower, though still above 3.5, suggesting acceptable translation quality for these languages.

4.3.3 Comparison Between Pairwise and Pointwise Score Transfer

We examine the relative effectiveness of pairwise versus pointwise judgment methods for transferring English scoring capabilities to multilingual settings. Based on the translation quality evaluations, we select two high-performing languages, Arabic and Spanish, for the study. Specifically, we translate 200 English text pairs into Arabic and 200 pairs into Spanish. Each dataset is then annotated by GPT-4o using both pairwise and pointwise scoring strategies. For pointwise annotation, GPT-4o assigns quality scores on a 1–10 scale. The scoring prompts of both methods explicitly instruct GPT-4o to evaluate based on content quality alone, irrespective of language and are detailed in Appendix C.3. Each text or pair is scored 20 times, and the average is used as the final score. Given identical content across different languages, the ideal scenario is that a consistent model and prompt should yield nearly identical scores, regardless of the language and score strategies.

As shown in Figure 6, pointwise scores exhibit greater variability across languages, particularly in the mid-quality range (scores between 3 and 6), despite relative stable assessments at the high and low ends. In contrast, pairwise judgments demonstrate high consistency across languages, with only minor pair-score discrepancies observed. These findings suggest that while translation quality is generally sufficient, minor translation biases can still influence pointwise ratings. The pairwise approach, however, proves more robust to such variation, supporting its suitability for reliably transferring English scoring capabilities to multilingual contexts.



(a) Arabic (Pointwise) (b) Spanish (Pointwise) (c) Arabic (Pairwise) (d) Spanish (Pairwise)

Figure 6: Scatter plots of average scores assigned by GPT-4o to Arabic and Spanish parallel data. Each point represents an average of 20 evaluations. Left: pointwise scoring. Right: pairwise scoring.

5 Conclusion

We have introduced MuRating, a unified and scalable approach to multilingual data selection that leverages high-quality English quality signals and transfers them across seventeen target languages. By first aggregating multiple English raters through a pairwise Bradley–Terry framework and then projecting those judgments via translation into monolingual, cross-lingual, and parallel text pairs, MuRating yields a single language-agnostic scorer capable of annotating trillions of tokens with minimal compute. When used to curate a 500 billion-token pretraining corpus (200 billion English + 300 billion multilingual), MuRating outperforms strong baselines—including uniform sampling (+50 % data), QuRater, AskLLM, and DCLM—on a suite of twelve English benchmarks and a diverse multilingual evaluation. Our experiments further demonstrate that pairwise translation supervision is more robust than pointwise scoring and cross-lingual and parallel-pair regularization greatly improve score consistency. While MuRating particularly boosts performance on knowledge-intensive and QA-style tasks, it underselects narrative-rich content, pointing to future avenues in genre-aware filtering and dynamic sampling schedules. We will release our prompts, code, and scored datasets to facilitate broader research into multilingual data quality and more equitable LLM development.

Limitations

Currently we've tried it on 17 tongues—there's plenty more to explore. Relying on GPT-4o brings its own quirks (and blind spots), and our taste for factual content means we don't give stories and creative writing their fair shake. We've only tested on a 1.2 B-parameter model, and scoring trillions of tokens still eats up a ton of GPU hours. In the future, we'll lean on better translations, expand to more languages, and add smarter sampling to capture a richer, more diverse mix of text.

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

[2] Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, 2024.

[3] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv:1911.11641*, 2019.

[4] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

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

[6] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[9] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2924–2936, 2019.

[10] 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. *arXiv preprint arXiv:1803.05457*, 2018.

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

[12] Ona De Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer Van Der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, et al. A new massive multilingual dataset for high-performance language technologies. *arXiv preprint arXiv:2403.14009*, 2024.

[13] Meng Fang, Yuan Li, and Trevor Cohn. Learning how to active learn: A deep reinforcement learning approach. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[14] 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. A framework for few-shot language model evaluation, 07 2024.

[15] Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

[16] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

[18] Wenhan Han, Yifan Zhang, Zhixun Chen, Binbin Liu, Haobin Lin, Bingni Zhang, Taifeng Wang, Mykola Pechenizkiy, Meng Fang, and Yin Zheng. Mubench: Assessment of multilingual capabilities of large language models across 61 languages, 2025.

[19] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2021.

[20] Tao Jiang, Xu Yuan, Yuan Chen, Ke Cheng, Liangmin Wang, Xiaofeng Chen, and Jianfeng Ma. Fuzzydedup: Secure fuzzy deduplication for cloud storage. *IEEE Transactions on Dependable and Secure Computing*, 20(3):2466–2483, 2022.

[21] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

[22] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[23] Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. Large language models only pass primary school exams in Indonesia: A comprehensive test on IndoMMLU. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 2023. Association for Computational Linguistics.

[24] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

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

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

[27] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Yichong Xu, Yujia Qin, Zihan Liu, Yiming Cui, and Yue Zhang. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics (ACL)*, 2024.

[28] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

[29] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*, 2020.

[30] Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024.

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

[32] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*, 2018.

[33] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*, 2016.

[34] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

[35] 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. *arXiv preprint arXiv:2309.09400*, 2023.

[36] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[37] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

[38] 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. *arXiv preprint arXiv:2306.01116*, 2023.

[39] Ru Peng, Kexin Yang, Yawen Zeng, Junyang Lin, Dayiheng Liu, and Junbo Zhao. Dataman: Data manager for pre-training large language models. *arXiv preprint arXiv:2502.19363*, 2025.

[40] Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*, 2020.

[41] Jirui Qi, Raquel Fernández, and Arianna Bisazza. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, 2023.

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

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

[44] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*, 2019.

[45] Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Zhengzhong Liu, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, et al. Slimpajama-dc: Understanding data combinations for llm training. *arXiv preprint arXiv:2309.10818*, 2023.

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

[47] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, 2024.

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

[49] Alexey Tikhonov and Max Ryabinin. It's all in the heads: Using attention heads as a baseline for cross-lingual transfer in commonsense reasoning. *arXiv preprint arXiv:2106.12066*, 2021.

[50] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

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

[52] Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, Yihong Chen, Raphael Tang, and Pontus Stenetorp. Multilingual pretraining using a large corpus machine-translated from a single source language. *arXiv preprint arXiv:2410.23956*, 2024.

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

[54] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin, editors, *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[55] Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. Qurating: Selecting high-quality data for training language models. *arXiv preprint arXiv:2402.09739*, 2024.

[56] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36:69798–69818, 2023.

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

[58] Liangyu Xu, Xuemiao Zhang, Feiyu Duan, Sirui Wang, Jingang Wang, and Xunliang Cai. Fire: Flexible integration of data quality ratings for effective pre-training. *arXiv preprint arXiv:2502.00761*, 2025.

[59] Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. To repeat or not to repeat: Insights from scaling llm under token-crisis. *Advances in Neural Information Processing Systems*, 36:59304–59322, 2023.

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

[61] 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*, pages 4791–4800, 2019.

A Details of MuRater Model

A.1 Different Annotation Method

GPT annotation We adopt the educational value prompt criteria from QuRating [55] as our annotation prompt for GPT-4o-08-06, as detailed below. This prompt is used to annotate a total of 300,000 document pairs. For each pair, we randomly extract a segment of n tokens—based on the LLaMA tokenizer [50]—where n is sampled from a uniform distribution $n \sim \text{Uniform}[256, 512]$ in 50% of cases, and fixed at 512 tokens otherwise. Annotation involves generating 20 predictions of either “A” or “B” per criterion and document pair (in either order). The total cost of dataset creation amounts to $9,740.

Pairwise Educational Value Prompt

Compare two text excerpts and choose the text which has more educational value, e.g., it includes clear explanations, step-by-step reasoning, or questions and answers.
Aspects that should NOT influence your judgement:
1. Which language the text is written in
2. The length of the text
3. The order in which the texts are presented
Note that the texts are cut off, so you have to infer their contexts. The texts might have similar quality, but you should still make a relative judgement and choose the label of the preferred text.
[Option label a] ... text a ...
[Option label b] ... text b ...
Now you have to choose between either label a or label b. Respond only with a single word.

Askllm We adopt the approach from [43] and use the following prompt to query Flan-T5-xxl [8] for annotation 300,000 document pairs.

Ask-LLM prompt

This is a pretraining .... datapoint.
Does the previous paragraph demarcated within ### and ### contain informative signal for pre-training a large-language model? An informative datapoint should be well-formatted, contain some usable knowledge of the world, and strictly NOT have any harmful, racist, sexist, etc. content.
OPTIONS:
- yes
- no

Fineweb and DCLM For these two data selection methods, we directly use the open-sourced model to annotate documents to obtain the scores.

A.2 Training details of MuRater and training accuracy

We adopt the XLM-RoBERTa architecture encoder model BGE-M3 [6] as the foundation of our multilingual rating model, MuRater, and fine-tune it by appending a linear regression head to the transformer output to predict quality scores. The fine-tuning process employs a confidence margin threshold of 50%, defined as $|p_A - p_B| = |2p_{B \succ A} - 1|$ for a prediction between text pairs (t_A, t_B) [55]. Fine-tuning is conducted over 3 epochs with a batch size of 512 and a learning rate of 2×10^{-5}. We set λ to 0.5. Performance on held-in and held-out sets is summarized in Table 3. Notably, BGE-M3 supports over 100 languages and leverages large-scale multilingual unsupervised data to learn a shared semantic space, making it particularly effective for multilingual and cross-lingual retrieval and rating tasks.

Evaluation Dataset	Confidence Margin	Accuracy
Held-in	50%	94.3%
	80%	97.2%
Held-out	50%	90.7%
	80%	93.1%

Table 3: Prediction accuracy of MuRater on held-in and held-out datasets under different confidence margins.

B Experiment Setup Details

B.1 Dataset

We apply the Fineweb and Fineweb-2 [37] as our training corpus. We employ NVIDIA's multilingual domain classifier[8] to label the domain distribution of our dataset. The Figures 7-11 depict the domain distributions before and after applying MuRater-based selection. The results indicate that MuRater consistently prioritizes knowledge-intensive domains, such as *People and Society*, *Health*, and *Science*. These domains are generally characterized by well-organized content and high informational density, which are advantageous for the pretraining of LLMs. Nevertheless, the selected domain distributions are not same across languages, primarily due to significant domain differences inherent in their respective source corpora.

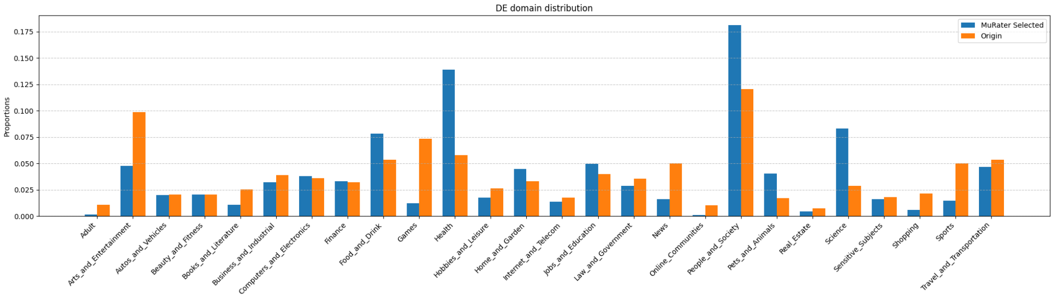


Figure 7: Domain distribution of German corpus

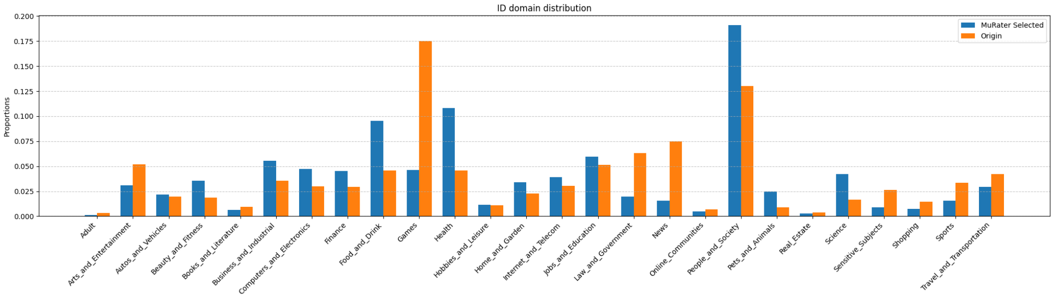


Figure 8: Domain distribution of France corpus

B.2 Baselines

We follow the same annotation procedure for the English datasets of QuRater, AskLLM, DCLM, and FineWeb-Edu as described in Appendix A. For QuRater-M, we apply the same prompting approach (also detailed in Appendix A) and instruct GPT-4o to annotate 300,000 multilingual pairs, focusing exclusively on content regardless of the language. We then fine-tune the multilingual QuRater baseline using both English and multilingual data, leveraging the BGE-M3 model [6] and the identical training hyperparameters outlined in Appendix A.

[8] `https://huggingface.co/nvidia/multilingual-domain-classifier`

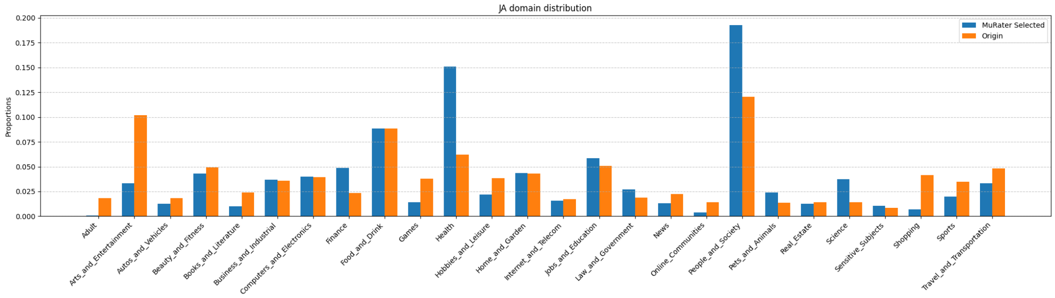


Figure 9: Domain distribution of Japanese corpus

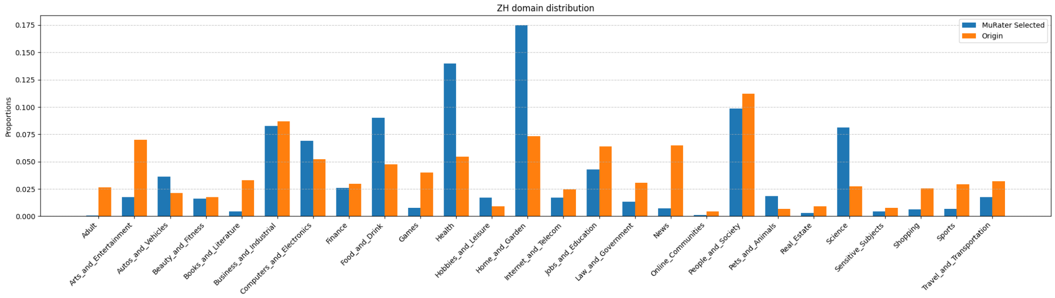


Figure 10: Domain distribution of Chinese corpus

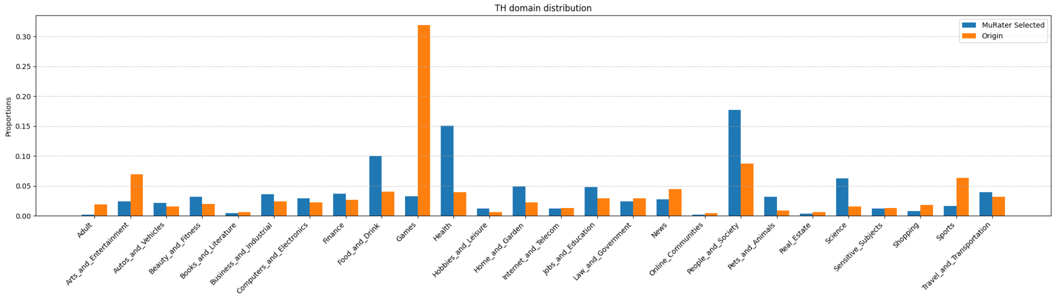


Figure 11: Domain distribution of Thai corpus

B.3 Model Architecture

We utilize a transformer architecture based on the LLaMA-2 model [50], configured to contain approximately 1.2 billion parameters. Models are randomly initialized before pretraining. The detailed information for the model configuration and training hyperparamters is shown in Table 4. We preprocess our training corpus to train a custom Byte-Pair Encoding (BPE) tokenizer using the BBPE algorithm, yielding a vocabulary of 250,000 tokens for use in our training experiments. The main experiments is conducted using 64 NVIDIA H100 GPUs, with an average runtime of approximately 70 hours per experiment.

B.4 Benchmarks

C Evaluation Benchmarks

All task evaluations are conducted using the `lm-evaluation-harness` framework [14]. For English in-context learning tasks, we use the following benchmakrs:

- **ARC-Easy and ARC-Challenge** [10] (25-shot): Multiple-choice science questions from grade school exams, assessing models' ability to apply scientific knowledge and reasoning.

Model configuration	Values
Attention head	16
Layers	24
Hiddent size	2048
Intermediate layer dimension	5504
maximum position embedding	4096
layer normalization epsilon	1×10^{-5}
Training Hyperparameters	**Values**
Batch size	3072
Sequence length	4096
Optimizer	AdamW
Learning rate	4.3×10^{-4}
Learning rate schedule	Cosine decay to 10% of inital value
Traning steps	Varied based on the total token budget
Precision	bf16(mxied-precision training)

Table 4: Model configuration and Training Hyperparameters for pretraining LLms

- **SciQ** [54] (0-shot): Crowdsourced multiple-choice science questions covering physics, chemistry, and biology, designed to evaluate scientific understanding.
- **LogiQA** [29] (0-shot): Logical reasoning questions derived from Chinese civil service exams, testing deductive reasoning capabilities.
- **TriviaQA** [21] (5-shot): Reading comprehension dataset with question-answer pairs authored by trivia enthusiasts, accompanied by evidence documents.
- **BoolQ** [9] (5-shot): Yes/no questions with associated passages, evaluating models' ability to answer naturally occurring questions.

For commonsense reasoning, we evaluate on:

- **HellaSwag** [61] (10-shot): Sentence completion tasks requiring commonsense inference to select the most plausible continuation.
- **PIQA** [3] (5-shot): Physical commonsense reasoning questions, focusing on everyday tasks and interactions.
- **OpenBookQA** [32] (10-shot): Multiple-choice questions based on elementary science facts, requiring both factual knowledge and reasoning.
- **WinoGrande** [44] (5-shot): Pronoun resolution tasks designed to test commonsense reasoning at scale.

Additionally, two knowledge-intensive tasks are evaluated:

- **Natural Questions (NQ)** [24] (5-shot): Real user questions paired with answers from Wikipedia, assessing open-domain question answering.
- **MMLU** [19] (5-shot): A benchmark covering 57 subjects across various domains, measuring multitask language understanding.

For evaluating translated benchmarks, we use the MuBench dataset [18] and conduct evaluations across 18 languages present in our training set. In the multilingual setting, we evaluate:

- **ARC-Easy and ARC-Challenge** (25-shot): Translated versions of the science question benchmarks, assessing cross-lingual reasoning.
- **HellaSwag** (10-shot): Evaluating commonsense reasoning in multiple languages through sentence completion tasks.
- **MMLU** (5-shot): Multilingual evaluation of multitask language understanding across diverse subjects.
- **StoryCloze** [33] (0-shot): Narrative understanding task where models choose the correct ending to a four-sentence story.

- **BMLAMA** [41] (0-shot): Multilingual factual knowledge probing dataset, assessing cross-lingual consistency in language models.
- **XCOPA** [40] (5-shot): Causal commonsense reasoning tasks translated into multiple languages, evaluating cross-lingual inference.
- **XNLI** [11] (5-shot): Cross-lingual natural language inference benchmark, testing entailment and contradiction detection.
- **XWinograd** [49] (5-shot): Multilingual pronoun resolution tasks, assessing commonsense reasoning across languages.
- **FLORES** [15] (5-shot): Multilingual machine translation benchmark, evaluating translation quality across diverse languages.
- **MMLU_L** (5-shot): A localized version of MMLU, focusing on both general knowledge and language-specific knowledge and reasoning tasks.

C.1 Translation

The translation prompts is

Translation Prompt

Please translate the following {lang} text into {lang2}. Your translations must convey all the content in the original text and cannot involve explanations or other unnecessary information. Please ensure that the translated text is natural for native speakers with correct grammar and proper word choices.
Your translation must also use exact terminology to provide accurate information even for the experts in the related fields.
The text is : {text}

We translate a total of 600,000 English document pairs evenly across 17 languages using the GPT-4o-08-06 model, with the overall translation cost amounting to $18,720.

C.2 Human translation quality evaluation

To assess the translation quality of GPT-4o outputs, we engaged professional human translators to evaluate a subset of the generated translations. Each language translation was reviewed by a single expert. Evaluators were compensated at a rate of $16 per hour, with each assessment session lasting approximately 4 hours. the annotation criteria for translation quality is shown below.

Annotation Criteria

5 points: The translation accurately reflects the meaning of the original text, is fluent, and contains no errors.
4 points: The translation generally reflects the meaning of the original text, with most sentences being fluent, but there are slight inaccuracies in the use of non-key terms or non-idiomatic phrases.
3 points: The translation conveys the general idea of the original text, but contains significant errors such as improper translation of key terms, incorrect word order, omissions, mistranslations, or untranslated segments.
2 points: The translation is largely incomprehensible or unfaithful to the original text, with serious errors including issues of order, logic, or severe grammatical mistakes.
1 point: The translation is completely incomprehensible or entirely unfaithful to the original text, or it fails to convey the original meaning entirely, being obscure and difficult to understand.
Please note that all sentences are excerpts from web content, so the last sentence of each segment, which may be unclear, is not considered in the evaluation.

We ensured adherence to ethical standards in our human annotation process:

- **Fair Compensation**: All annotators received compensation at or above the minimum wage standards of their respective regions.
- **Informed Consent**: Annotators were provided with clear instructions and information about the annotation tasks. Participation was voluntary, and informed consent was obtained prior to their involvement.
- **Institutional Review**: Our study underwent review and received approval from the Institutional Review Board (IRB) at our institution, ensuring that the research met ethical standards for studies involving human participants.
- **Transparency**: Detailed information regarding the annotation are included in the supplementary materials to promote transparency and reproducibility.

C.3 Pointwise Score

The pointwise scoring prompt is provided below. We instruct GPT-4o to evaluate each text 10 times, then compute the average of these scores to determine the final rating. The scoring range is from `grade_min = 1` to `grade_max = 10`.

> Pointwise prompt evaluation for educational value
>
> I need to rate a text excerpt on a scale of {grade_min} to {grade_max} (inclusive) based on its educational value, e.g., it includes clear explanations, step-by-step reasoning, or questions and answers.
> Aspects that should NOT influence your judgement: 1. Which language the text is written in 2. The length of the text
> Note that the text is cut off, so you have to infer its context.
> [Text] ... {text} ...
> Now assign a number grade between {grade_min} to {grade_max} (inclusive). Respond only with a single digit. The score for the quality of the text is:

D Detailed Results

D.1 English Detailed Results

Table 5 presents the detailed performance of various selection methods across individual downstream tasks. Our method consistently outperforms others on most tasks, with notable improvements on ARC, HellaSwag, and MMLU.

Table 5: Detailed performance of differnt selection method over all downstream tasks with all values in percentages and per-benchmark maximum highlighted in bold.

Data Selection Method	ARC_Challenge	ARC_Easy	BoolQ	HellaSwag	LogiQA	MMLU	NQ	OpenBookQA	PIQA	TriviaQA	WinoGrande	SciQ	Average
Uniform (*+50% data*)	35.24	66.50	64.46	62.90	28.88	32.85	**7.87**	37.00	75.73	27.00	**60.62**	85.40	48.70
Askllm	36.60	67.63	59.76	63.33	26.57	32.89	7.53	35.60	76.82	26.55	57.85	82.70	47.82
DCLM	40.44	73.78	64.07	62.42	28.73	35.42	9.31	37.40	76.06	28.01	60.06	87.00	50.23
FineWeb_Edu	40.10	72.39	**64.62**	59.06	26.88	36.01	7.98	**38.20**	74.27	**29.05**	58.41	86.90	49.49
QuRater	40.27	72.14	61.93	62.38	28.88	35.26	5.68	38.60	75.63	15.74	57.70	85.80	48.33
MuRater	**43.77**	**75.84**	64.28	**65.06**	**30.11**	**37.24**	7.81	**38.20**	**77.04**	28.69	59.51	**87.20**	**51.23**

D.2 Multilingual Detailed Results

We display the detailed results of each benchmark and each language below.

Method	AR	DE	EN	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TA	TH	TR	VI	ZH
Uniform	42.21	53.07	65.57	56.40	53.66	52.02	52.86	49.07	44.44	45.62	51.43	54.17	50.67	37.88	39.44	46.72	48.44	55.47
QuRater-M	52.69	62.25	72.94	65.74	63.59	61.32	62.71	57.62	53.58	54.29	60.44	63.51	59.34	42.85	43.90	55.72	54.67	63.72
MuRater(M)	52.19	62.79	72.85	66.54	63.85	63.30	62.58	58.00	53.96	55.89	**61.70**	63.47	59.85	43.35	**45.75**	56.40	56.19	63.76
MuRater(E)	**52.82**	**63.22**	**73.91**	**67.55**	**63.97**	**63.68**	**63.80**	**58.88**	**54.59**	**56.78**	61.62	**65.24**	**60.98**	**44.53**	45.50	**57.28**	**57.03**	**65.11**

Table 6: Detailed per-language performance on across **ARC-Easy**. Bold indicates the best result for each language.

Method	AR	DE	EN	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TA	TH	TR	VI	ZH
Uniform	27.82	30.03	32.25	30.12	30.03	28.92	30.03	28.92	26.54	29.01	28.33	30.55	29.61	24.83	28.24	28.16	28.58	28.92
QuRater-M	30.55	35.58	41.89	36.95	35.92	34.90	37.20	33.87	32.59	34.56	34.13	36.60	35.32	**28.16**	28.75	32.34	32.59	36.18
MuRater(M)	30.20	**36.95**	41.13	**39.25**	35.75	35.07	35.49	34.39	33.36	32.51	34.56	37.71	35.84	27.99	29.78	33.45	**33.70**	**36.35**
MuRater(E)	**31.91**	36.09	**42.06**	39.08	**37.29**	**36.18**	**38.57**	**35.67**	**33.45**	**36.01**	**34.81**	**39.25**	**37.20**	27.13	**30.20**	**35.07**	31.48	35.84

Table 7: Detailed per-language performance on across **ARC-Challenge**. Bold indicates the best result for each language.

Method	AR	DE	EN	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TA	TH	TR	VI	ZH
Uniform	39.52	47.50	60.48	51.46	51.37	47.01	49.52	42.09	38.53	43.34	48.36	50.17	45.76	36.68	35.62	40.05	43.95	46.59
QuRater-M	41.83	49.71	61.61	54.05	54.48	49.95	51.87	43.88	**40.63**	45.10	50.20	52.71	**48.92**	37.93	37.46	42.53	**46.33**	47.50
MuRater(M)	41.65	49.62	**62.46**	54.00	54.62	49.94	51.89	43.53	40.32	45.60	50.08	52.63	48.03	37.41	37.10	42.15	45.33	47.23
MuRater(E)	**42.17**	**50.23**	62.30	**54.84**	**55.13**	**50.36**	**52.13**	**44.39**	40.48	**46.16**	**50.89**	**53.55**	48.53	37.69	37.30	**42.62**	46.28	**48.06**

Table 8: Detailed per-language performance on across **HellaSwag**. Bold indicates the best result for each language.

Method	AR	DE	EN	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TH	TL	TR	VI	ZH
Uniform	0.2628	0.2769	0.2968	0.2782	0.2810	0.2787	0.2741	0.2777	0.2748	0.2791	0.2772	0.2817	0.2708	0.2701	0.2743	0.2742	0.2799	0.2824
QuRater-M	0.2747	0.2949	0.3180	0.2935	0.2975	0.2988	0.2915	0.2911	0.2852	0.2880	0.2979	0.2953	0.2893	**0.2812**	0.2821	0.2872	0.2915	0.2947
MuRater(M)	0.2727	0.2957	0.3235	0.2908	0.3018	**0.3000**	0.2944	0.2909	**0.2919**	0.2877	0.2968	0.2997	0.2907	0.2797	0.2812	0.2874	0.2944	0.2914
MuRater(E)	**0.2765**	**0.3033**	**0.3206**	**0.2983**	**0.3010**	0.2989	**0.2905**	**0.2936**	0.2871	**0.2925**	**0.2976**	**0.2988**	**0.2886**	0.2813	**0.2850**	**0.2868**	**0.2967**	**0.2949**

Table 9: Detailed per-language performance on across **MMLU**. Bold indicates the best result for each language.

Method	AR	DE	EN	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TH	TL	TR	VI	ZH
Uniform	0.6161	**0.7237**	**0.7570**	0.7221	**0.7237**	0.6974	**0.6950**	**0.6718**	**0.6463**	**0.6881**	**0.7074**	**0.7214**	0.7090	0.6502	**0.5967**	**0.6238**	**0.6865**	**0.7136**
QuRater-M	0.6014	0.6912	0.7291	0.6927	0.7005	0.6703	0.6726	0.6633	0.5983	0.6471	0.6780	0.6912	0.6757	0.6269	0.5797	0.5875	0.6610	0.6881
MuRater(M)	0.6037	0.6989	0.7314	0.7074	0.7059	0.6989	0.6803	0.6649	0.6246	0.6656	0.6943	0.7098	0.6974	**0.6393**	0.5820	0.6029	0.6811	0.6865
MuRater(E)	**0.6231**	0.7082	0.7307	**0.7059**	0.7012	**0.6950**	0.6834	0.6811	0.6416	0.6610	0.6927	0.6981	**0.7144**	0.6517	0.5967	0.6122	0.6850	0.6981

Table 10: Detailed per-language performance on across **StoryCloze**. Bold indicates the best result for each language.

Method	AR	DE	EN	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TH	TL	TR	VI	ZH
Uniform	0.3128	0.4530	0.5148	0.4531	0.4563	0.3860	0.4422	0.4082	0.2764	0.3536	0.4001	0.4026	0.3391	0.2862	0.4702	0.3063	0.4294	0.4387
QuRater-M	0.3850	0.5540	0.5838	0.5075	0.4860	0.4757	0.5076	0.4317	0.3388	0.4629	0.5186	0.4835	0.3846	0.3431	0.5040	0.3971	0.4934	0.3931
MuRater(M)	0.4039	0.5660	0.6331	0.5725	0.5582	0.5544	0.5563	0.4353	0.3389	0.5364	0.5404	0.5194	0.4350	0.3679	0.5519	0.4393	0.5643	0.4500
MuRater(E)	**0.4451**	**0.6062**	**0.6380**	**0.5919**	**0.5549**	**0.5898**	**0.5828**	**0.4749**	**0.3920**	**0.5703**	**0.5933**	**0.5578**	**0.4646**	**0.3828**	**0.5615**	**0.4576**	**0.6034**	**0.4669**

Table 11: Detailed per-language performance on across **BMLAMA**. Bold indicates the best result for each language.

Method	ID	IT	TH	TR	VI	ZH
Uniform	68.20	66.60	57.20	58.80	**70.60**	66.20
QuRater-M	65.20	65.00	56.00	58.40	67.40	65.80
MuRater(M)	67.80	67.20	**58.20**	58.80	69.00	65.60
MuRater(E)	**69.00**	**68.20**	57.20	**60.20**	70.20	**69.60**

Table 12: Detailed per-language performance on across **XCOPA**. Bold indicates the best result for each language.

Method	AR	DE	EN	ES	FR	RU	TH	TR	VI	ZH
Uniform	35.90	46.47	47.67	45.74	46.14	43.29	38.35	39.60	39.56	40.60
QuRater-M	**37.11**	47.63	49.60	**47.71**	49.32	46.99	37.87	43.78	**41.93**	41.93
MuRater(M)	35.74	44.34	46.79	44.50	47.15	44.14	**38.39**	39.60	38.80	41.24
MuRater(E)	34.86	**48.84**	**51.49**	46.55	**49.40**	**47.39**	37.27	**43.94**	41.77	**43.13**

Table 13: Detailed per-language performance on across **XNLI**. Bold indicates the best result for each language.

Method	EN	FR	JP	PT	RU	ZH
Uniform	**83.70**	**69.88**	**67.78**	69.96	62.86	**72.02**
QuRater-M	77.12	66.27	66.21	66.54	60.32	63.49
MuRater(M)	78.54	65.06	67.47	67.30	62.86	67.86
MuRater(E)	80.22	**69.88**	66.32	**71.48**	**65.71**	68.25

Table 14: Detailed per-language performance on **XWinograd**. Bold indicates the best result for each language.

Method	AR	DE	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TH	TL	TR	VI	ZH
Uniform	35.03	53.27	48.27	57.70	57.95	47.76	21.81	18.99	53.94	48.94	58.22	44.17	28.60	40.04	37.94	48.97	19.82
QuRater-M	36.60	52.70	48.28	58.74	59.62	48.17	23.59	19.52	54.23	48.20	59.03	45.09	29.94	42.05	39.73	48.65	19.97
MuRater(M)	**37.84**	53.65	**48.92**	**59.45**	60.17	48.85	**24.01**	**21.26**	**54.33**	49.51	**60.30**	**46.70**	**30.78**	41.83	40.11	**50.89**	19.98
MuRater(E)	37.80	**53.87**	48.30	58.85	**60.20**	**49.39**	23.73	20.99	54.03	**49.52**	60.05	46.14	29.84	**42.49**	**40.64**	50.79	**20.40**

(a) Translation from English (EN TO ML)

Method	AR	DE	ES	FR	ID	IT	JA	KO	MS	NL	PT	RU	TH	TL	TR	VI	ZH
Uniform	52.14	**61.41**	**54.46**	61.71	57.93	55.28	**42.48**	41.73	56.63	**54.41**	64.62	54.96	45.81	**49.27**	45.40	52.24	**46.10**
QuRater-M	51.14	60.40	53.63	61.47	57.81	55.68	42.13	41.04	55.82	53.77	64.34	53.30	44.44	47.39	46.58	51.23	44.62
MuRater(M)	52.39	60.86	54.26	**62.64**	57.77	**56.21**	42.47	**42.47**	**56.71**	54.26	**64.80**	**55.00**	45.65	48.84	46.17	**52.53**	45.40
MuRater(E)	**52.63**	60.93	54.03	61.72	**57.98**	56.00	42.25	41.48	56.12	54.23	64.59	54.55	**46.01**	47.97	**47.03**	52.04	45.31

(b) Translation to English (ML TO EN)

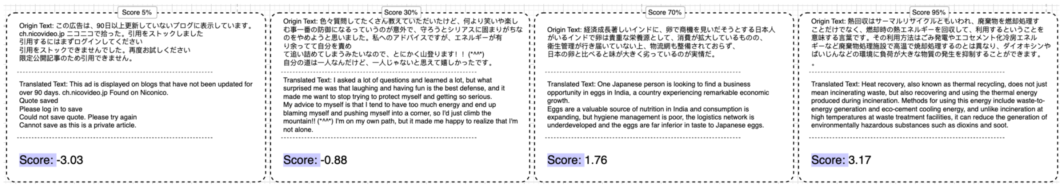
Table 15: Detailed per-language performance on **FLORES**. Bold indicates the best result for each language.

Method	AMMLU	CMMLU	INDOMMLU	JMMLU	VLMU
Uniform	0.2594	0.3175	0.3235	0.3079	0.2909
QuRater-M	0.2659	0.3398	0.3278	0.3197	0.2898
MuRater(M)	0.2713	**0.3467**	0.3441	0.3304	0.3005
MuRater(E)	**0.2714**	0.3404	**0.3489**	**0.3323**	**0.3048**

Table 16: Detailed per-language performance on across **MMLU-L**. Bold indicates the best result per column.

E Case Study

We present examples from various languages exhibiting a range of quality scores. The results demonstrate that texts with higher scores tend to be more fluent and contain richer educational content, particularly in domains such as health and science. Moreover, for texts with comparable scores, the quality remains consistent across different languages. This suggests that our MuRater model evaluates text quality in a language-agnostic manner, relying solely on the content rather than the language in which it is written.



Score 5%

Origin Text: この広告は、90日以上更新していないブログに表示しています。
ch.nicovideo.jp ニコニコで拾った。引用をストックしました
引用するにはまずログインしてください
引用をストックできませんでした。再度お試しください
限定公開記事のため引用できません。

Translated Text: This ad is displayed on blogs that have not been updated for over 90 days. ch.nicovideo.jp Found on Niconico.
Quote saved
Please log in to save
Could not save quote. Please try again
Cannot save as this is a private article.

Score: -3.03

Score 30%

Origin Text: 色々質問してたくさん教えていただいたけど、何より笑いや楽しむ事一番の防御になるっていうのが意外で、守ろうとシリアスに固まりがちなのをやめようと思いました。私へのアドバイスですが、エネルギーが有り余ってて自分を責め
て追い詰めてしまうみたいなので、とにかく山登ります！！ (*^^*)
自分の道は一人なんだけど、一人じゃないと思えて嬉しかったです。

Translated Text: I asked a lot of questions and learned a lot, but what surprised me was that laughing and having fun is the best defense, and it made me want to stop trying to protect myself and getting so serious.
My advice to myself is that I tend to have too much energy and end up blaming myself and pushing myself into a corner, so I'd just climb the mountain!! (*^^*) I'm on my own path, but it made me happy to realize that I'm not alone.

Score: -0.88

Score 70%

Origin Text: 経済成長著しいインドに、卵で商機を見いだそうとする日本人がいるインドで卵は貴重な栄養源として、消費が拡大しているものの、
衛生管理が行き届いていない上、物流網も整備されておらず、
日本の卵と比べると味が大きく劣っているのが実情だ。

Translated Text: One Japanese person is looking to find a business opportunity in eggs in India, a country experiencing remarkable economic growth.
Eggs are a valuable source of nutrition in India and consumption is expanding, but hygiene management is poor, the logistics network is underdeveloped and the eggs are far inferior in taste to Japanese eggs.

Score: 1.76

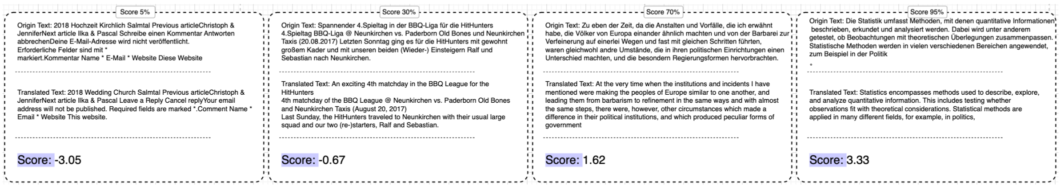
Score 95%

Origin Text: 熱回収はサーマルリサイクルともいわれ、廃棄物を燃却処理することだけでなく、燃却時の熱エネルギーを回収して、利用するということを意味する言葉です。その利用方法はごみ発電やエコセメント化冷房エネルギーなど廃棄物処理施設で高温で焼却処理するのとは異なり、ダイオキシンやばいじんなどの環境に負荷が大きな物質の発生を抑制することができます。
。

Translated Text: Heat recovery, also known as thermal recycling, does not just mean incinerating waste, but also recovering and using the thermal energy produced during incineration. Methods for using this energy include waste-to-energy generation and eco-cement cooling energy, and unlike incineration at high temperatures at waste treatment facilities, it can reduce the generation of environmentally hazardous substances such as dioxins and soot.

Score: 3.17

Figure 12: Sampled training examples of **Japanese** with quality ratings at different score range



Score 5%

Origin Text: 2018 Hochzeit Kirchlich Salmtal Previous articleChristoph &
JenniferNext article Ilka & Pascal Schreibe einen Kommentar Antworten
abbrechenDeine E-Mail-Adresse wird nicht veröffentlicht.
Erforderliche Felder sind mit *
markiert.Kommentar Name * E-Mail * Website Diese Website

Translated Text: 2018 Wedding Church Salmtal Previous articleChristoph &
JenniferNext article Ilka & Pascal Leave a Reply Cancel replyYour email
address will not be published. Required fields are marked *.Comment Name *
Email * Website This website.

Score: -3.05

Score 30%

Origin Text: Spannender 4.Spieltag in der BBQ-Liga für die HitHunters
4.Spieltag BBQ-Liga @ Neunkirchen vs. Paderborn Old Bones und Neunkirchen
Taxis (20.08.2017) Letzten Sonntag ging es für die HitHunters mit gewohnt
großem Kader und mit unseren beiden (Wieder-) Einsteigern Ralf und
Sebastian nach Neunkirchen.

Translated Text: An exciting 4th matchday in the BBQ League for the
HitHunters
4th matchday of the BBQ League @ Neunkirchen vs. Paderborn Old Bones
and Neunkirchen Taxis (August 20, 2017)
Last Sunday, the HitHunters traveled to Neunkirchen with their usual large
squad and our two (re-)starters, Ralf and Sebastian.

Score: -0.67

Score 70%

Origin Text: Zu eben der Zeit, da die Anstalten und Vorfälle, die ich erwähnt
habe, die Völker von Europa einander ähnlich machten und von der Barbarei zur
Verfeinerung auf einerlei Wegen und fast mit gleichen Schritten führten,
waren gleichwohl andre Umstände, die in ihren politischen Einrichtungen einen
Unterschied machten, und die besondern Regierungsformen hervorbrachten.

Translated Text: At the very time when the institutions and incidents I have
mentioned were making the peoples of Europe similar to one another, and
leading them from barbarism to refinement in the same ways and with almost
the same steps, there were, however, other circumstances which made a
difference in their political institutions, and which produced peculiar forms of
government

Score: 1.62

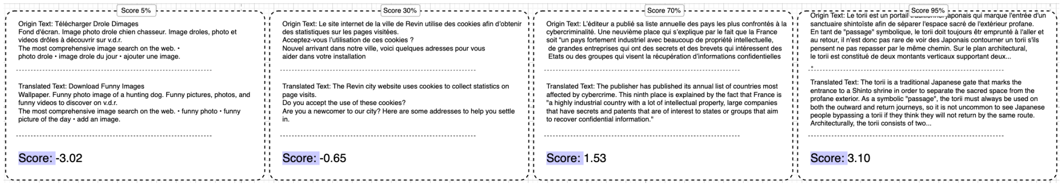
Score 95%

Origin Text: Die Statistik umfasst Methoden, mit denen quantitative Informationen
beschrieben, erkundet und analysiert werden. Dabei wird unter anderem
getestet, ob Beobachtungen mit theoretischen Überlegungen zusammenpassen.
Statistische Methoden werden in vielen verschiedenen Bereichen angewendet,
zum Beispiel in der Politik
。

Translated Text: Statistics encompasses methods used to describe, explore,
and analyze quantitative information. This includes testing whether
observations fit with theoretical considerations. Statistical methods are
applied in many different fields, for example, in politics,

Score: 3.33

Figure 13: Sampled training examples of **German** with quality ratings at different score range



Score 5%

Origin Text: Télécharger Drole Dimages
Fond d'écran. Image photo drole chien chasseur. Image droles, photo et
videos drôles à découvrir sur v.d.r.
The most comprehensive image search on the web. •
photo drole • image drole du jour • ajouter une image.

Translated Text: Download Funny Images
Wallpaper. Funny photo image of a hunting dog. Funny pictures, photos, and
funny videos to discover on v.d.r.
The most comprehensive image search on the web. • funny photo • funny
picture of the day • add an image.

Score: -3.02

Score 30%

Origin Text: Le site internet de la ville de Revin utilise des cookies afin d'obtenir
des statistiques sur les pages visitées.
Acceptez-vous l'utilisation de ces cookies ?
Nouvel arrivant dans notre ville, voici quelques adresses pour vous
aider dans votre installation

Translated Text: The Revin city website uses cookies to collect statistics on
page visits.
Do you accept the use of these cookies?
Are you a newcomer to our city? Here are some addresses to help you settle
in.

Score: -0.65

Score 70%

Origin Text: L'éditeur a publié sa liste annuelle des pays les plus confrontés à la
cybercriminalité. Une neuvième place qui s'explique par le fait que la France
soit "un pays fortement industriel avec beaucoup de propriété intellectuelle,
de grandes entreprises qui ont des secrets et des brevets qui intéressent des
Etats ou des groupes qui visent la récupération d'informations confidentielles

Translated Text: The publisher has published its annual list of countries most
affected by cybercrime. This ninth place is explained by the fact that France is
"a highly industrial country with a lot of intellectual property, large companies
that have secrets and patents that are of interest to states or groups that aim
to recover confidential information."

Score: 1.53

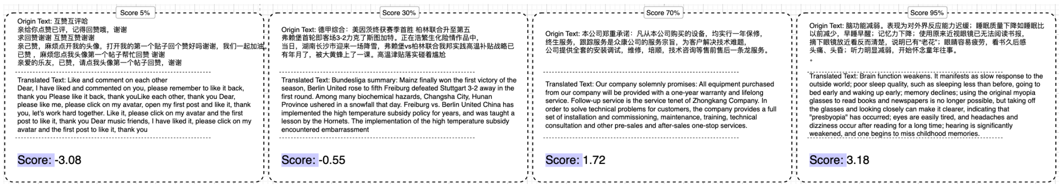
Score 95%

Origin Text: Le torii est un portail traditionnel japonais qui marque l'entrée d'un
sanctuaire shintoïste afin de séparer l'espace sacré de l'extérieur profane.
En tant de "passage" symbolique, le torii doit toujours êtr emprunté à l'aller et
au retour, il n'est donc pas rare de voir des Japonais contourner un torii s'ils
pensent ne pas repasser par le même chemin. Sur le plan architectural,
le torii est constitué de deux montants verticaux supportant deux...
。

Translated Text: The torii is a traditional Japanese gate that marks the
entrance to a Shinto shrine in order to separate the sacred space from the
profane exterior. As a symbolic "passage", the torii must always be used on
both the outward and return journeys, so it is not uncommon to see Japanese
people bypassing a torii if they think they will not return by the same route.
Architecturally, the torii consists of two...

Score: 3.10

Figure 14: Sampled training examples of **France** with quality ratings at different score range



Score 5%

Origin Text: 互赞互评哈
亲给你点赞已评，记得回赞哦，谢谢
求回赞谢谢 互赞互赞谢谢
亲己赞，麻烦点开我的头像，打开我的第一个贴子回个赞好吗谢谢，我们一起加油。
已赞，麻烦您点我头像第一个帖子帮忙回赞 谢谢
亲爱的乐友，已赞，请点我头像第一个帖子回赞，谢谢

Translated Text: Like and comment on each other
Dear, I have liked and commented on you, please remember to like it back,
thank you Please like it back, thank youLike each other, thank you Dear,
please like me, please click on my avatar, open my first post and like it, thank
you, let's work hard together. Like it, please click on my avatar and the first
post to like it, thank you Dear music friends, I have liked it, please click on my
avatar and the first post to like it, thank you

Score: -3.08

Score 30%

Origin Text: 德甲综合：美因茨终获赛季首胜 柏林联合升至第五
弗赖堡首轮即客场3-2力克了斯图加特。正在浩繁生化险情作品中，
当日，湖南长沙市迎来一场降雪，弗赖堡vs柏林联合我邦实践高温补贴战略已
有年月了，被大黄蜂上了一课。高温津贴落实碰着尴尬

Translated Text: Bundesliga summary: Mainz finally won the first victory of the
season, Berlin United rose to fifth Freiburg defeated Stuttgart 3-2 away in the
first round. Among many biochemical hazards, Changsha City, Hunan
Province ushered in a snowfall that day. Freiburg vs. Berlin United China has
implemented the high temperature subsidy policy for years, and was taught a
lesson by the Hornets. The implementation of the high temperature subsidy
encountered embarrassment

Score: -0.55

Score 70%

Origin Text: 本公司郑重承诺：凡从本公司购买的设备，均实行一年保修，
终生服务，跟踪服务是众康公司的服务宗旨，为客户解决技术难题，
公司提供全套的安装调试，维修，培顺，技术咨询等售前售后一条龙服务。

Translated Text: Our company solemnly promises: All equipment purchased
from our company will be provided with a one-year warranty and lifelong
service. Follow-up service is the service tenet of Zhongkang Company. In
order to solve technical problems for customers, the company provides a full
set of installation and commissioning, maintenance, training, technical
consultation and other pre-sales and after-sales one-stop services.

Score: 1.72

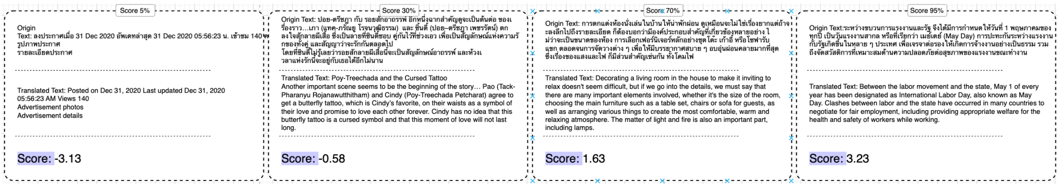
Score 95%

Origin Text: 脑功能减弱。表现为对外界反应能力迟缓；睡眠质量下降如睡眠比
以前减少，早睡早醒；记忆力下降；使用原来近视眼镜已无法阅读书报，
摘下眼镜放近看反而清楚，说明已有"老花"；眼睛容易疲劳，看书久后感
头痛、头昏；听力明显减弱，开始怀念童年往事。
。

Translated Text: Brain function weakens. It manifests as slow response to the
outside world; poor sleep quality, such as sleeping less than before, going to
bed early and waking up early; memory declines; using the original myopia
glasses to read books and newspapers is no longer possible, but taking off
the glasses and looking closely can make it clearer, indicating that
"presbyopia" has occurred; eyes are easily tired, and headaches and
dizziness occur after reading for a long time; hearing is significantly
weakened, and one begins to miss childhood memories.

Score: 3.18

Figure 15: Sampled training examples of **Chinese** with quality ratings at different score range



Score 5%

Origin
Text: ลงประกาศเมื่อ 31 Dec 2020 อัพเดทล่าสุด 31 Dec 2020 05:56:23 น. เข้าชม 140 ค
รูปภาพประกาศ
รายละเอียดประกาศ

Translated Text: Posted on Dec 31, 2020 Last updated Dec 31, 2020
05:56:23 AM Views 140
Advertisement photos
Advertisement details

Score: -3.13

Score 30%

Origin Text: ปอย-ตรีชฎา กับ รอยสักอาถรรพ์ อีกหนึ่งฉากสำคัญดูจะเป็นต้นต่อ ของเ
รื่องราว…เภา (แทค-ภรัณยู โรจนวุฒิธรรม) และ ซินดี้ (ปอย–ตรีชฎา เพชรรัตน์) ตก
ลงใจสักลายผีเสื้อ ซึ่งเป็นลายที่ซินดี้ชอบ คู่กันไว้ที่ช่วงเอว เพื่อเป็นสัญลักษณ์แห่งความรั
กของทั้งคู่ และสัญญาว่าจะรักกันตลอดไป
โดยที่ซินดี้ไม่รู้เลยว่ารอยสักลายผีเสื้อนี้จะเป็นสัญลักษณ์อาถรรพ์ และห้วงเ
วลาแห่งรักนี้จะอยู่กับเธอได้อีกไม่นาน

Translated Text: Poy-Treechada and the Cursed Tattoo
Another important scene seems to be the beginning of the story… Pao (Tack-
Pharanyu Rojanawutthitham) and Cindy (Poy-Treechada Petcharat) agree to
get a butterfly tattoo, which is Cindy's favorite, on their waists as a symbol of
their love and promise to love each other forever. Cindy has no idea that this
butterfly tattoo is a cursed symbol and that this moment of love will not last
long.

Score: -0.58

Score 70%

Origin Text: การตกแต่งห้องนั่งเล่นในบ้านให้น่าพักผ่อน ดูเหมือนจะไม่ใช่เรื่องยากแต่ถ้าจ
ะลงลึกไปถึงรายละเอียด ก็ต้องบอกว่ามีองค์ประกอบสำคัญที่เกี่ยวข้องหลายอย่าง ไ
ม่ว่าจะเป็นขนาดของห้อง การเลือกเฟอร์นิเจอร์หลักอย่างชุดโต๊ะ เก้าอี้ หรือโซฟารับ
แขก ตลอดจนการจัดวางต่าง ๆ เพื่อให้มีบรรยากาศสบาย ๆ อบอุ่นผ่อนคลายมากที่สุด
ซึ่งเรื่องของแสงและไฟ ก็มีส่วนสำคัญเช่นกัน ทั้งโคมไฟ

Translated Text: Decorating a living room in the house to make it inviting to
relax doesn't seem difficult, but if we go into the details, we must say that
there are many important elements involved, whether it's the size of the room,
choosing the main furniture such as a table set, chairs or sofa for guests, as
well as arranging various things to create the most comfortable, warm and
relaxing atmosphere. The matter of light and fire is also an important part,
including lamps.

Score: 1.63

Score 95%

Origin Text:ระหว่างขบวนการแรงงานและรัฐ จึงได้มีการกำหนดให้วันที่ 1 พฤษภาคมของ
ทุกปี เป็นวันแรงงานสากล หรือที่เรียกว่า เมย์เดย์ (May Day) การปะทะกันระหว่างแรงงาน
กับรัฐเกิดขึ้นในหลาย ๆ ประเทศ เพื่อเจรจาต่อรองให้เกิดการจ้างงานอย่างเป็นธรรม รวม
ถึงจัดสวัสดิการที่เหมาะสมด้านความปลอดภัยต่อสุขภาพของแรงงานขณะทำงาน

Translated Text: Between the labor movement and the state, May 1 of every
year has been designated as International Labor Day, also known as May
Day. Clashes between labor and the state have occurred in many countries to
negotiate for fair employment, including providing appropriate welfare for the
health and safety of workers while working.

Score: 3.23

Figure 16: Sampled training examples of **Thai** with quality ratings at different score range