

A Details of MuRater Model

A.1 Different Annotation Method

GPT annotation We adopt the educational value prompt criteria from QuRating [53] 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 [48]—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 [41] 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) [53]. Fine-tuning is conducted over 2 epochs with a batch size of 512 and a learning rate of 2×10^{-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 curate 16 recent snapshots from Common Crawl⁷, specifically: CC-MAIN-2021-39, 2021-43, 2021-49, 2022-05, 2022-21, 2022-27, 2022-33, 2022-40, 2022-49, 2023-06, 2023-14, 2023-23, 2023-40, 2023-50, 2024-10, and 2024-18. These snapshots are processed using deduplication and heuristic filtering pipelines adapted from SlimPajama [44] and FineWeb [28]. This preprocessing yields a multilingual corpus containing approximately 3 trillion tokens. A similar methodology is then employed to extract 1.5 trillion English tokens by filtering the most recent 9 Common Crawl snapshots.

We employ NVIDIA’s multilingual domain classifier⁸ 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.

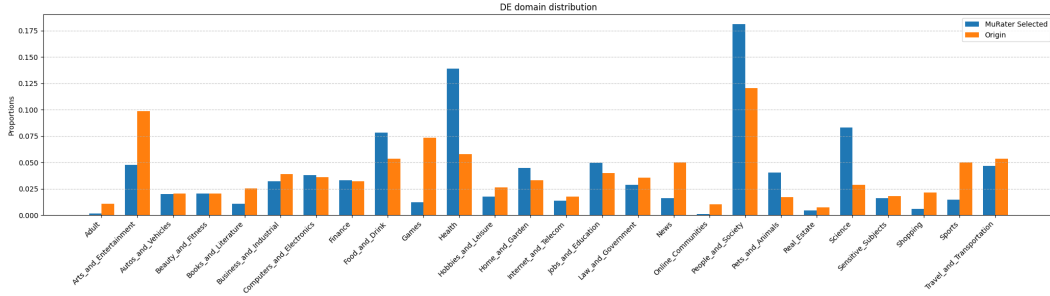


Figure 7: Domain distribution of German corpus

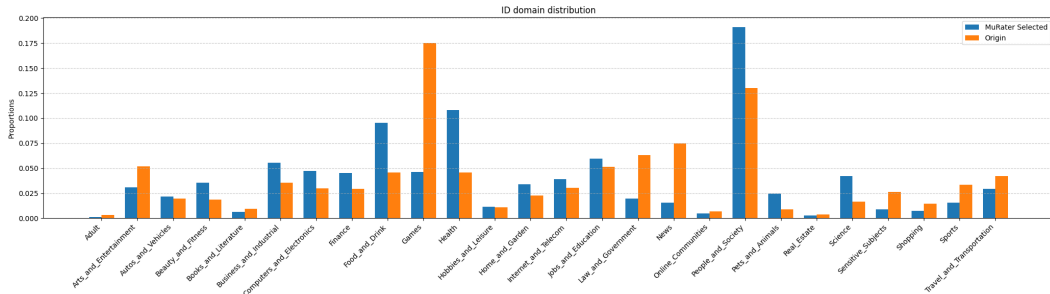


Figure 8: Domain distribution of France corpus

⁷<https://commoncrawl.org/>

⁸<https://huggingface.co/nvidia/multilingual-domain-classifier>

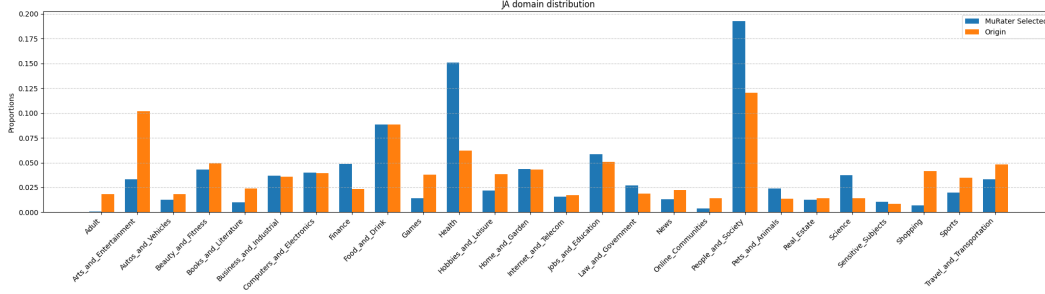


Figure 9: Domain distribution of Japanese corpus

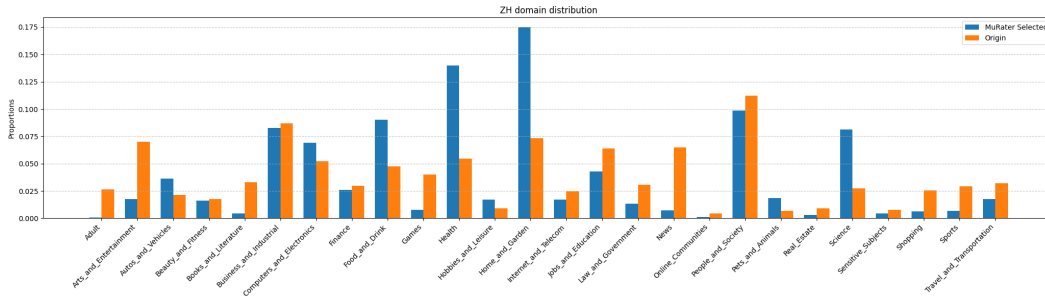


Figure 10: Domain distribution of Chinese corpus

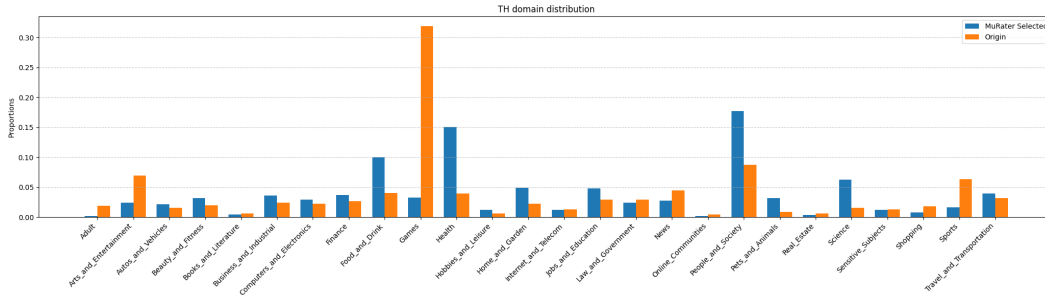


Figure 11: Domain distribution of Thai corpus

939 B.2 Baselines

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

946 B.3 Model Architecture

947 We utilize a transformer architecture based on the LLaMA-2 model [48], configured to contain
 948 approximately 1.2 billion parameters. Models are randomly initialized before pretraining. The
 949 detailed information for the model configuration and training hyperparameters is shown in Table 4. We
 950 preprocess our training corpus to train a custom Byte-Pair Encoding (BPE) tokenizer using the BBPE
 951 algorithm, yielding a vocabulary of 250,000 tokens for use in our training experiments. The main
 952 experiments is conducted using 64 NVIDIA H100 GPUs, with an average runtime of approximately
 953 50 hours per experiment.

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

B.4 Benchmarks

C Evaluation Benchmarks

All task evaluations are conducted using the `lm-evaluation-harness` framework [13]. 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.
- **SciQ** [52] (0-shot): Crowdsourced multiple-choice science questions covering physics, chemistry, and biology, designed to evaluate scientific understanding.
- **LogiQA** [27] (0-shot): Logical reasoning questions derived from Chinese civil service exams, testing deductive reasoning capabilities.
- **TriviaQA** [19] (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** [59] (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** [30] (10-shot): Multiple-choice questions based on elementary science facts, requiring both factual knowledge and reasoning.
- **WinoGrande** [42] (5-shot): Pronoun resolution tasks designed to test commonsense reasoning at scale.

Additionally, two knowledge-intensive tasks are evaluated:

- **Natural Questions (NQ)** [22] (5-shot): Real user questions paired with answers from Wikipedia, assessing open-domain question answering.
- **MMLU** [17] (5-shot): A benchmark covering 57 subjects across various domains, measuring multitask language understanding.

For evaluating translated benchmarks, we use the MuBench dataset⁹ and conduct evaluations across 18 languages present in our training set. In the multilingual setting, we evaluate:

⁹<https://huggingface.co/datasets/aialt/MuBench>

- 984 • **ARC-Easy and ARC-Challenge** (25-shot): Translated versions of the science question
985 benchmarks, assessing cross-lingual reasoning.
- 986 • **HellaSwag** (10-shot): Evaluating commonsense reasoning in multiple languages through
987 sentence completion tasks.
- 988 • **MMLU** (5-shot): Multilingual evaluation of multitask language understanding across diverse
989 subjects.
- 990 • **StoryCloze** [31] (0-shot): Narrative understanding task where models choose the correct
991 ending to a four-sentence story.
- 992 • **BMLAMA** [39] (0-shot): Multilingual factual knowledge probing dataset, assessing cross-
993 lingual consistency in language models.
- 994 • **XCOPA** [38] (5-shot): Causal commonsense reasoning tasks translated into multiple lan-
995 guages, evaluating cross-lingual inference.
- 996 • **XNLI** [11] (5-shot): Cross-lingual natural language inference benchmark, testing entailment
997 and contradiction detection.
- 998 • **XWinograd** [47] (5-shot): Multilingual pronoun resolution tasks, assessing commonsense
999 reasoning across languages.
- 1000 • **FLORES** [14] (5-shot): Multilingual machine translation benchmark, evaluating translation
1001 quality across diverse languages.
- 1002 • **MMLU_L** (5-shot): A localized version of MMLU, focusing on both general knowledge
1003 and language-specific knowledge and reasoning tasks.

1004 C.1 Translation

1005 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}

1006

1007 We translate a total of 600,000 English document pairs evenly across 17 languages using the GPT-4o-
1008 08-06 model, with the overall translation cost amounting to \$18,720.

1009 C.2 Human translation quality evaluation

1010 To assess the translation quality of GPT-4o outputs, we engaged professional human translators to
1011 evaluate a subset of the generated translations. Each language translation was reviewed by a single
1012 expert. Evaluators were compensated at a rate of \$16 per hour, with each assessment session lasting
1013 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.

1014

1015 In alignment with the NeurIPS Code of Ethics and the Paper Checklist Guidelines, we ensured
1016 adherence to ethical standards in our human annotation process:

- 1017 • **Fair Compensation:** All annotators received compensation at or above the minimum wage
1018 standards of their respective regions, as mandated by NeurIPS guidelines.
- 1019 • **Informed Consent:** Annotators were provided with clear instructions and information about
1020 the annotation tasks. Participation was voluntary, and informed consent was obtained prior
1021 to their involvement.
- 1022 • **Institutional Review:** Our study underwent review and received approval from the Institu-
1023 tional Review Board (IRB) at our institution, ensuring that the research met ethical standards
1024 for studies involving human participants.
- 1025 • **Transparency:** Detailed information regarding the annotation are included in the supple-
1026 mentary materials to promote transparency and reproducibility.

1027 C.3 Pointwise Score

1028 The pointwise scoring prompt is provided below. We instruct GPT-4o to evaluate each text 10 times,
1029 then compute the average of these scores to determine the final rating. The scoring range is from
1030 `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:

1031

1032 D Detailed Results

1033 D.1 English Detailed Results

1034 Table 5 presents the detailed performance of various selection methods across individual downstream
1035 tasks. Our method consistently outperforms others on most tasks, with notable improvements on
1036 ARC, HellaSwag, and MMLU.

Table 5: Detailed performance of different 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

1037 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.

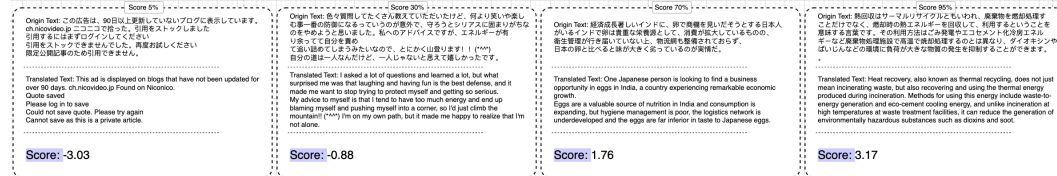


Figure 12: Sampled training examples of **Japanese** with quality ratings at different score range

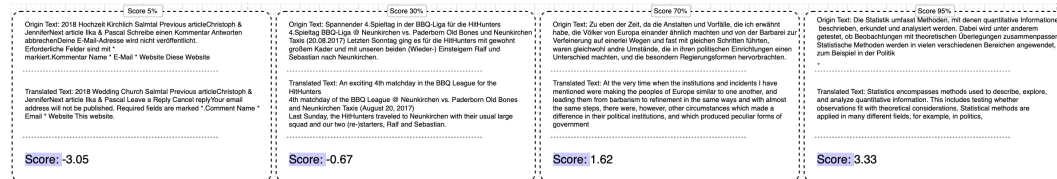


Figure 13: Sampled training examples of **German** with quality ratings at different score range

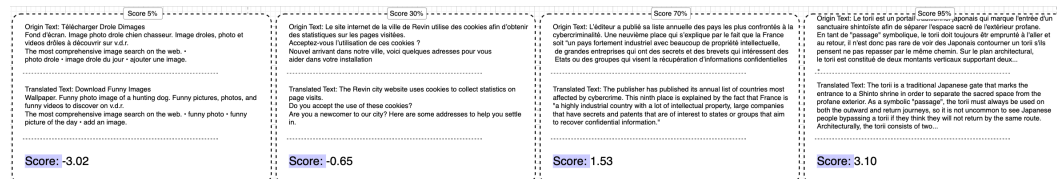


Figure 14: Sampled training examples of **France** with quality ratings at different score range



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

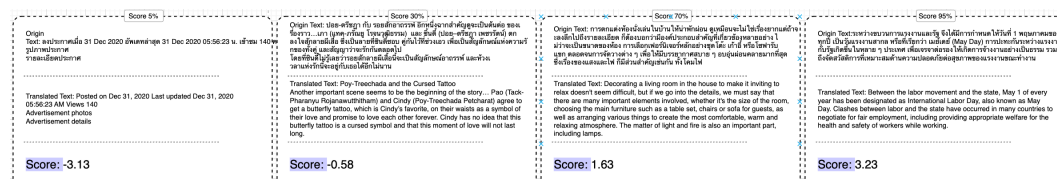


Figure 16: Sampled training examples of **Thai** with quality ratings at different score range