

# MTQ-Eval: Multilingual Text Quality Evaluation for Language Models

Rhitabrat Pokharel and Ameeta Agrawal

Department of Computer Science  
Portland State University, USA  
{pokharel, ameeta}@pdx.edu

## Abstract

The use of large language models (LLMs) for evaluating outputs is becoming an increasingly effective and scalable approach. However, it remains uncertain whether this capability extends beyond task-specific evaluations to more general assessments of text quality, particularly in multilingual contexts. In this study, we introduce, MTQ-Eval, a novel framework for multilingual text quality evaluation that learns from examples of both high- and low-quality texts, adjusting its internal representations. To develop MTQ-Eval, we first automatically generate text quality preference data and then use it to train open-source base LLMs to align with ratings of high- and low-quality text. Our comprehensive evaluation across 115 languages demonstrates the improved performance of the proposed model. Upon further analysis, we find that this enhanced evaluation capability also leads to notable improvements in downstream tasks.

## 1 Introduction

Recent advances in LLMs have significantly enhanced their performance across various tasks, including text generation. While human evaluation remains the gold standard for assessing the quality of the generated text, it is costly, time-consuming, and particularly challenging to scale across languages. As a result, researchers have turned to LLMs themselves as automated evaluators for tasks such as translation, dialogue assessment, and essay scoring (Zheng et al., 2023; Gilardi et al., 2023; Wang et al., 2023; Kocmi and Federmann, 2023; Ferron et al., 2023; Stahl et al., 2024; Nam et al., 2024).

Evaluating text quality is essential to ensure outputs are not only task-specifically accurate but also coherent, fluent, and natural to a native speaker. However, there is still no standardized framework for assessing general quality of text using LLMs, especially in multilingual settings where linguistic

diversity and data availability present significant challenges (Gala et al., 2023; Bagheri Nezhad and Agrawal, 2024), and as LLM capabilities in multilingual contexts remain inconsistent (Hada et al., 2024b; Agrawal et al., 2024).

Prompt-based text quality evaluation faces challenges in multilingual scenarios. For instance, while LLM-as-judge methods such as G-Eval (Liu et al., 2023) can assess aspects like fluency and coherence, they are mainly effective in few-shot settings (Doostmohammadi et al., 2024), which may not be practical for languages with limited data. Additionally, the scalability of prompt-based methods may depend on the model’s proficiency in understanding the target language. Furthermore, existing methods prioritize task-specific evaluation (Hada et al., 2024b; Hu et al., 2024; Bagheri Nezhad et al., 2025) over general text quality. Complicating matters further, the concept of text quality remains broadly referenced but lacks a formal and universally accepted definition. In the literature, it has been discussed in various contexts such as natural language generation (NLG) quality (Hu et al., 2024), task quality, or content quality (Hada et al., 2024b,a).

Consider the example in Table 1 where LLMs such as GPT-4o and Llama 3.1 struggle even in English to assess the logical coherence of a passage where the first passage of superior quality should have received a higher score (4 or 5 on a scale of 1 to 5). This raises a natural question: *Can a model reliably assess text quality across multiple languages, including low-resource languages?* This question leads to another compelling hypothesis: *Does a model’s ability to effectively evaluate text quality translate to improved performance in other downstream tasks such as text classification or summarization?*

To answer these questions, we introduce the Multilingual Text Quality Evaluation framework (MTQ-Eval), summarized in Figure 1. Instead of relying

Passage	GPT-4o	Llama-3.1	MTQ-Eval
<b>Original Text:</b> Red tide is caused by a higher than normal concentration of <i>Karenia brevis</i> , a naturally-occurring single-celled marine organism. Natural factors can intersect to produce ideal conditions, allowing this algae to increase in number dramatically. The algae produces a neurotoxin that can disable nerves in both humans and fish. Fish often die because of the high concentrations of the toxin in the waters. Humans can be affected by breathing affected water taken into the air by wind and waves.	4	3	4
<b>Word Shuffled:</b> Red tide is caused by a higher than normal concentration of <i>Karenia brevis</i> , a naturally-occurring single-celled marine organism. Natural factors can intersect <b>humans</b> produce ideal conditions, allowing this algae to increase in number dramatically. The algae produces a neurotoxin that can disable nerves in both <b>to</b> and fish. Fish often die because of the high concentrations of the toxin in the waters. Humans <b>by</b> be affected <b>can</b> breathing <b>waves</b> . water taken into the air by <b>and wind affected</b>	2	2	1

Table 1: Text quality ratings from three different models for the original and degraded text reveal that the scores for the original text are questionable for the llama-3.1 model but these are improved when using MTQ-Eval.

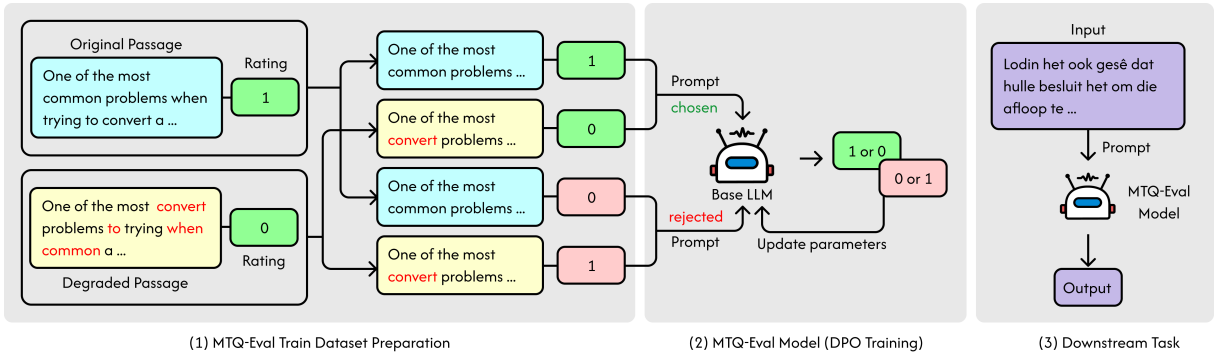


Figure 1: Overview of the MTQ-Eval (1) dataset creation, (2) model training, and (3) evaluation.

on prompt-based evaluations that require language proficiency, we propose training a model to distinguish between high- and low-quality text. A major bottleneck in multilingual evaluation is the lack of good quality annotated datasets across languages. To overcome this, we develop an efficient and scalable synthetic data generation process that does not rely on human annotations, enabling the model to generalize across 115 languages.

Our work makes the following contributions:

- We propose a novel framework<sup>1</sup> using Direct Preference Optimization (DPO) (Rafailov et al., 2023) to align models with text quality preference data, improving text evaluation in a wide range of languages.
- We automatically generate a dataset of high- and low-quality text samples across 115 languages without requiring human annotations.
- We conduct extensive multilingual evaluations comparing our approach against baseline

<sup>1</sup>The code is publicly available at <https://github.com/PortNLP/MTQ-Eval> and the models at <https://huggingface.co/PortNLP>

methods and also investigate whether proficiency in evaluating text quality maintains or improves performance in downstream tasks.

## 2 MTQ-Eval

We consider text quality to encompass the following dimensions: **coherence** (the logical flow and connectivity between sentences and ideas in the text as a whole), **fluency** (the smoothness and grammatical accuracy), **simplicity** (the ease of understanding), and **linguistic acceptability** (the naturalness and overall appropriateness), informed by prior research (Hu et al., 2024; Hada et al., 2024b).

### 2.1 Dataset Creation

To train LLMs to distinguish between high- and low-quality content across multiple languages, we need appropriate training data of preferences.

**High quality text** As base data, we leverage the high-quality text passages of Belebele<sup>2</sup> dataset

<sup>2</sup><https://huggingface.co/datasets/facebook/belebele>

(Bandarkar et al., 2024), which is built upon Flores-200 (Costa-jussà et al., 2022). We select this dataset for its human-translated content in 100+ languages and a well-distributed spread of languages across both high and low resource levels, with a minimum of 43 languages represented in each level. High-resource (HR) languages are categorized under levels 3, 4, and 5, while low-resource (LR) languages correspond to levels 0, 1, and 2, as defined by Joshi et al. (2020). Besides, it is a parallel dataset that facilitates fair comparisons across languages.

**Low quality text** Previous research has employed perturbation techniques such as reversing logical structures, altering verb inflections, and shuffling sentences. These methods are mostly feasible only for high-resource languages. Implementing them across more than hundred languages, most of them low-resource, is challenging due to the lack of robust NLP tools such as sentence tokenizers, connective detectors, and verb detectors.

Informed by prior work, we explore a controlled method for degrading text quality – word shuffling (Kallini et al., 2024; Hu et al., 2024). This straightforward technique, generally applicable across most languages, randomly rearranges a few words within a passage, disrupting the grammatical structure and semantic flow, creating noticeably distorted, low-quality text. For word shuffling, we tokenized text by spaces and randomly swapped three to six words within each sample to introduce enough distortion. Too much distortion is also not good for model training so that the degraded texts do not drift too far from the normal data. While useful, we acknowledge that this process may have varied impacts across languages, especially those with flexible word order.

Out of the 122 languages in Belebele, we do not consider 7 languages e.g., Basque and Japanese where word order plays lesser role and spaces do not function as reliable delimiters. This results in a dataset of 115 languages. From each language, we randomly selected 20 samples, yielding 2300 *normal* passage samples and 2300 *degraded* passages, for a total of 4,600 passages.

**Quality annotations** As obtaining human annotations at scale for over 100 languages is infeasible, an alternative approach is to assign binary values to both normal and degraded text – a score of 1 to the normal text, while a score of 0 to the corresponding degraded text (i.e., word shuffled text), as our model training (described in the next section)

is formulated around relative ranking between two options, where one text is explicitly preferred over another. A small-scale manual review of the low-quality data across a handful of languages familiar to the authors confirmed its adequacy.

## 2.2 Model Training

To align the model with text quality judgments, we use Direct Preference Optimization (DPO) which has shown to be simpler and more stable than reinforcement learning from human feedback (RLHF) (Rafailov et al., 2023). Moreover, unlike RLHF, DPO does not require explicit reward modeling, making it computationally efficient and easier to scale. DPO has been successfully used in various model aligning tasks like multilingual alignment (Lai et al., 2023; Wu et al., 2024; Pokharel et al., 2025), user intent alignment (Tunstall et al., 2024) and diffusion model alignment (Wallace et al., 2024). In this work, we explore this method to align a model to prefer high-quality text over low-quality text in multilingual settings.

DPO fine-tunes the model by increasing the likelihood of preferred responses while reducing the probability of less preferred ones. The model directly optimizes for relative preference ranking through a simple classification objective. The optimization follows a log-ratio loss function, which is defined as:

$$L_{\text{DPO}} = -\log \frac{\exp(\pi_{\theta}(x_{\text{good}}))}{\exp(\pi_{\theta}(x_{\text{good}})) + \exp(\pi_{\theta}(x_{\text{bad}}))}$$

where  $x_{\text{good}}$  represents a high-quality text sample,  $x_{\text{bad}}$  is a low-quality text sample, and  $\pi_{\theta}$  denotes the policy network (the LLM being fine-tuned).

This loss formulation encourages the model to rank the better text higher rather than simply classifying quality in isolation. By training on direct comparisons rather than absolute scores, DPO provides a robust framework for learning text preferences across diverse languages and linguistic structures.

A DPO method requires three key components from the dataset:

- *prompt* (the instruction given to the model),
- *chosen response* (the preferred high-quality text sample), and
- *rejected response* (the less preferred low-quality text sample).

To construct *prompts*, we ask the model to evaluate and rate a given passage. The prompt template is based on a combination of methodologies

Please act as an impartial judge and evaluate the text quality of the provided passage. Text quality of a passage is defined by how well it maintains the following aspects.

- (1) Coherence - logical flow and connectivity between sentences and ideas in the text.
- (2) Fluency - smoothness and naturalness of individual sentences.
- (3) Simplicity - how easy it is to understand the passage.
- (4) Linguistic Acceptability - if the text sounds natural and correct to a native speaker.

Provide a binary rating, "0" for low quality or "1" for high quality, strictly following this format: "[[0]]" or "[[1]]". Do not provide an explanation.

Passage: {passage}

Figure 2: Prompt used to obtain text quality ratings during DPO.

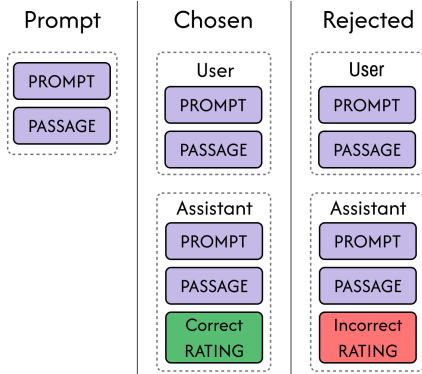


Figure 3: The training dataset format for DPO training.

from G-Eval (Liu et al., 2023) and LLM-as-a-Judge (Zheng et al., 2023), and is illustrated in Figure 2. For the *chosen* response, we provide the original passage along with its expected rating, while in the *rejected* response, we provide the same passage but with an incorrect rating. For instance, for original high-quality passages, a chosen rating is 1 and a rejected rating is 0. Similarly, for low-quality passages, we assign a chosen rating of 0 and a rejected rating of 1. This ensures a clear contrast between the ratings of normal passages and degraded texts. A simplified version of the data preparation process is illustrated in Figure 3 with detailed examples of each component provided in Appendix A.1.

### 3 Evaluation Setup

We conduct a comprehensive evaluation of our proposed model across various experimental settings as shown in Table 2. This section details the models, datasets, and evaluation metrics used in our study.

#### 3.1 Models and Baselines

We use Llama 3.1 Instruct 8B<sup>3</sup> (llama) and Aya Expanse 8B<sup>4</sup> (aya) as base models due to their

<sup>3</sup><https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct>

<sup>4</sup><https://huggingface.co/CohereForAI/aya-expense-8b>

Task	Train Data	Model
<b>Text-quality</b>		
Belebele	Train split	SFT model for Belebele
Belebele	Train split	Using DPO to create MTQ-Eval
MELA	Train split	SFT model for MELA
<b>Downstream</b>		
Sentiment Analysis (MMS)	Train split	SFT model for MMS
Summarization (XLSum)	Train split	SFT model for XLSum

Table 2: Overview of tasks, corresponding training data, and models used in our experiments.

extensive multilingual support: llama supports 8 languages, while aya performs well in 23 languages. During experiments, we compare MTQ-Eval (llama and aya) against base models using prompt-based inference and supervised fine-tuned (SFT) versions of these models. We train multiple SFT models specific to each task/dataset using 20 labeled samples per language from the corresponding dataset. For fine-tuning all models, we apply LoRA (Hu et al., 2022) with these configurations – alpha value of 128, dropout rate of 0.05, rank of 64, and one epoch of training with a batch size of 2 and a learning rate of 5e-7.

#### 3.2 Datasets and Metrics

We evaluate our models using two text quality evaluation datasets and two downstream task datasets.

##### 3.2.1 Text Quality Evaluation

**Linguistic Acceptability** We use the Multilingual Evaluation of Linguistic Acceptability (MELA) dataset (Zhang et al., 2024), which provides human ratings for linguistic acceptability across 10 languages (9 high-resource, 1 low-resource). This dataset assesses whether a sentence is syntactically acceptable to native speakers. We use the original



CoLA-like prompt (Warstadt et al., 2019) (detailed in Appendix A.2). We randomly sample 20 examples per language (total 200 samples) and report the results in terms of Matthew’s Correlation Coefficient (MCC).

### Text Quality Evaluation using Belebele Passages

While MELA provides human ratings, it is limited to only 10 languages. Scaling such judgments to 100+ languages is infeasible, so we construct an alternative test dataset using the methodology described in Section 2.1. We extract a different set of passages from Belebele and create a test set of 2,300 high-quality samples and 2,300 low-quality samples (20 samples each across 115 languages) resulting in a total of 4,600 test instances. The distribution includes 43 HR languages, 61 LR languages, and a few uncategorizable ones. This dataset also covers a diverse range of language families and scripts. During inference, we use the same prompt shown in Figure 2, with the addition of `Your Answer:` at the end. We report results using MCC along with Kullback-Leibler (KL) divergence which measures how distinguishable the ratings of normal text are compared to degraded text, as well as F1 scores.

### 3.3 Downstream Task Evaluation Datasets and Metrics

To assess whether improvements in text quality evaluation enhance performance in other NLP tasks, we evaluate models on multilingual text classification and text generation.

**Sentiment Analysis** We use the Massively Multilingual Corpus of Sentiment (MMS) dataset (Augustyniak et al., 2023) which includes 28 languages spanning six language families. The dataset is labeled with sentiment labels: positive, neutral, and negative. We select 20 samples per language, resulting in a total of 560 samples and report the results in terms of F1 scores. The prompt is based on LLM-as-a-Judge concept (Zheng et al., 2023) (details in Appendix A.2).

**Summarization** We use the XL-Sum dataset (Hasan et al., 2021) which contains article-summary pairs sourced from the BBC across 45 languages, ranging from low to high resource. We sample 20 articles from each language (total 900 samples) and generate summaries using a prompt adapted from (Chowdhery et al., 2023) (details in Appendix A.2). We evaluate the results using G-Eval (Liu et al.,

Lang.	11ama		aya	
	SFT	MTQ-Eval	SFT	MTQ-Eval
ara	<b>0.61</b>	0.00	0.31	<b>0.73</b>
deu	<b>0.00</b>	-0.25	<b>0.52</b>	0.31
eng	0.23	<b>0.35</b>	0.50	<b>0.52</b>
spa	0.00	<b>0.42</b>	0.23	<b>0.25</b>
fra	0.14	<b>0.25</b>	<b>0.31</b>	0.22
isl	<b>0.00</b>	-0.23	-0.23	<b>0.33</b>
ita	<b>0.33</b>	0.23	0.00	<b>0.42</b>
jpn	-0.14	<b>0.25</b>	<b>0.23</b>	0.20
rus	0.23	<b>0.33</b>	0.23	<b>0.42</b>
zho	0.33	<b>0.42</b>	0.00	<b>0.50</b>
Avg.	0.17	<b>0.18</b>	0.21	<b>0.39</b>

Table 3: MCC results on MELA dataset.

	MCC $\uparrow$	KL Div. $\uparrow$	F1 $\uparrow$
11ama base	0.18	0.06	0.56
11ama SFT	0.17	0.07	0.57
11ama MTQ-Eval	<b>0.24</b>	<b>0.12</b>	<b>0.59</b>
aya base	0.14	0.04	0.44
aya SFT	-0.23	0.10	0.35
aya MTQ-Eval	<b>0.26</b>	<b>0.15</b>	<b>0.60</b>

Table 4: MCC, KL Divergence, and F1 scores on Belebele text quality evaluation dataset.

2023) across three dimensions: coherence, consistency, and fluency.

## 4 Results of Text Quality Evaluation

In this section, we discuss the results of assessing multilingual text quality.

### 4.1 Performance on MELA dataset

Table 3 presents the MCC scores for text quality evaluation on the MELA dataset across ten languages using both 11ama and aya models, evaluated under SFT and MTQ-Eval paradigms. Across both models, the MTQ-Eval approach generally improves performance over SFT, particularly for aya, which achieves an average MCC of 0.39— a substantial improvement over SFT (0.21). 11ama sees a smaller but still positive gain.

For individual languages, 11ama MTQ-Eval is particularly beneficial for Spanish and Chinese, while aya MTQ-Eval shows strong gains in Arabic and Chinese. Interestingly, even English benefits from MTQ-Eval in both models. Overall, these results suggest that MTQ-Eval enhances multilingual text quality evaluation across most languages.

## 4.2 Performance on Belebele text quality dataset

Encouraged by these results, we extend our evaluation to a test set created using the Belebele dataset. From the average results covering 115 languages presented in Table 4, we observe that across both models, MTQ-Eval consistently outperforms both the base and SFT versions. The MCC scores show that MTQ-Eval correlates better with expected ratings (with an improvement in MCC from 0.18 to 0.24 for llama and 0.14 to 0.26 for aya), while the KL divergence results further confirm MTQ-Eval’s effectiveness in distinguishing between high- and low-quality text. The stark underperformance of aya SFT highlights the robustness of preference-based training over standard supervised fine-tuning. The F1 score results (the ground truth for normal text is 1, and for degraded text, it is 0), also support that the MTQ-Eval models are effectively distinguishing between normal and degraded text.

**High- vs. Low-Resource Languages** Regarding performance for high- and low-resource languages, the MCC scores in Table 5 show that MTQ-Eval generally shows improved performance for both models, particularly in high-resource languages. The gains in low-resource languages are less pronounced, reflecting the general challenge of low-resource modeling.

A further analysis of these results across a more granular breakdown of resource levels (0-5) is shown in Table 6. In low-resource settings (levels 0-2), llama MTQ-Eval achieves an average F1 score of 0.55, slightly outperforming aya (0.50). However, the gap widens in high-resource settings (levels 3-5) with aya significantly outperforming llama, achieving an average F1 score of 0.77 as compared to 0.65 for llama. This indicates that aya benefits more from MTQ-Eval for higher-resource languages.

**Supported Languages** Our next analysis specifically focuses on languages supported by llama (8 languages) and aya (23 languages). Figure 4 shows that for half of the model-supported languages, MTQ-Eval outperforms SFT under llama. However, aya shows a much stronger contrast between MTQ-Eval and SFT, with MTQ-Eval showing substantial improvements in all languages.

**Language Families** Next, we examine the performance across language families. The results presented in Table 7 show that llama MTQ-Eval

	MCC $\uparrow$		F1 $\uparrow$	
	LR	HR	LR	HR
llama base	0.11	0.29	0.52	0.61
llama SFT	0.11	0.29	0.54	0.63
llama MTQ-Eval	<b>0.16</b>	<b>0.38</b>	<b>0.55</b>	<b>0.64</b>
aya base	<b>0.12</b>	0.19	0.43	0.42
aya SFT	-0.06	-0.43	0.42	0.28
aya MTQ-Eval	0.07	<b>0.50</b>	<b>0.50</b>	<b>0.75</b>

Table 5: MCC and F1 scores on Belebele text quality evaluation dataset for high-resource (HR) and low-resource (LR) languages.

Res. Level	MTQ-Eval llama	MTQ-Eval aya
0	0.58	0.50
1	0.55	0.51
2	0.52	0.49
LR Avg.	0.55	0.50
3	0.61	0.69
4	0.68	0.82
5	0.67	0.82
HR Avg.	0.65	0.77

Table 6: F1 scores by resource levels on Belebele text quality evaluation dataset.

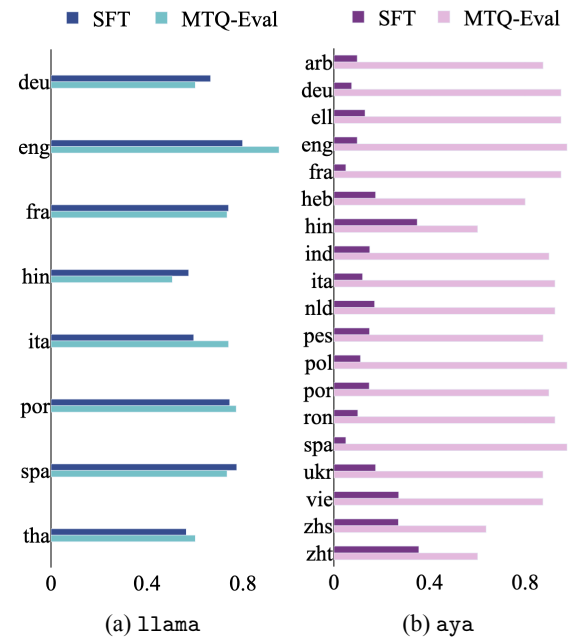


Figure 4: F1 scores of supported languages

performs best in Creole and Austronesian languages, while aya MTQ-Eval does well in Indo-European and Austro-Asiatic languages. Both models struggle with Nilo-Saharan languages, indicating a common weakness in underrepresented linguistic groups.

Lang. Family	MTQ-Eval 11ama	Lang. Family	MTQ-Eval aya
Creole	0.67	Indo-European	0.72
Austronesian	0.65	Austro-Asiatic	0.70
Turkic	0.63	Kartvelian	0.61
Indo-European	0.62	Afro-Asiatic	0.59
Uralic	0.61	Sino-Tibetan	0.57
Indo-Aryan	0.60	Austronesian	0.56
Afro-Asiatic	0.57	Dravidian	0.53
Tupian	0.55	Uralic	0.52
Niger-Congo	0.54	Kra-Dai	0.51
Austro-Asiatic	0.54	Turkic	0.48
Sino-Tibetan	0.53	Mongolic	0.46
Mongolic	0.52	Niger-Congo	0.34
Kra-Dai	0.52	Tupian	0.33
Kartvelian	0.39	Creole	0.33
Nilo-Saharan	0.37	Indo-Aryan	0.33
Dravidian	0.36	Nilo-Saharan	0.33

Table 7: F1 scores across language families on Belebele text quality evaluation dataset.

Table 8 presents the top and bottom performing languages (full results in Appendix A.3). For 11ama MTQ-Eval, the top 20 performing languages included both high-resource (60%) and low-resource (40%) languages, with 65% being Indo-European and 95% using the Latin script. In contrast, the bottom 20 consisted of 80% low-resource languages, with several diverse scripts. For aya MTQ-Eval, the top 20 was exclusively high-resource languages, with 80% Indo-European and 75% Latin script. The bottom 20 was dominated by low-resource languages, 65% of which were Niger-Congo, and 85% used the Latin script. Interestingly, right-to-left script languages (Persian, Arabic) show strong performance in both models. Overall, both models perform well in high-resource European languages, while showing weaknesses in South Asian scripts and African languages.

**Scripts** Table 9 presents the percentage improvement by script with MTQ-Eval over base models. While 11ama MTQ-Eval benefits Khmer, Malayalam, and Tibetan scripts, aya MTQ-Eval shows the largest improvement for Hebrew, Simplified Chinese (Hans), and Tamil scripts. However, 11ama MTQ-Eval negatively impacts some scripts (Orya, Tamil, and Bengali). Overall, aya benefits more from MTQ-eval across a wider range of scripts as compared to 11ama.

## 5 Downstream Task Performance

Having verified MTQ-Eval’s effectiveness in text quality assessment, we now evaluate whether its benefits extend to downstream tasks like sentiment analysis and summarization.

### 5.1 Sentiment Analysis

Table 10 shows the F1 scores on the MMS dataset. When trained on 11ama, the average F1 score of 0.51 with MTQ-Eval and 0.49 with SFT demonstrates that our method performs slightly better. aya MTQ-Eval sees similar improvements with average F1 score of 0.52, compared to 0.48 with SFT. Recall that MTQ-Eval is trained on synthetic text quality data while SFT is trained on task-specific labeled data. While the gains are modest, the results suggest that models trained to assess text quality also develop stronger representations, which help with classification tasks. More importantly, a key advantage of our method is that it does not require training with *task-specific* labeled data. Instead, with only a small set of synthetically generated text quality samples, it can boost downstream performance. This is particularly valuable in multilingual settings, where obtaining sufficient training data for low-resource languages is often challenging.

### 5.2 Summarization

Table 11 presents the mean scores of evaluating summary quality across high- and low-resource languages of XL-Sum dataset. We observe that MTQ-Eval performs as well as or better than SFT across all three evaluation dimensions, and particularly in low-resource categories ((+0.03 in Llama, +0.01 in Aya-Expanse)). This demonstrates that MTQ-Eval naturally enhances the ability to generate high-quality content, without explicit training on task-specific data.

We also analyzed English (HR) and Nepali (LR) summaries to explore behavior across language settings in Table 12. MTQ-Eval shows only slight improvements in English, where SFT already generates strong summaries. In contrast, for Nepali, where SFT outputs are noisy, MTQ-Eval produces coherent summaries. This highlights its greater impact in LR languages compared to HR ones.

## 6 Related Work

**LLM-as-an-Evaluator** The use of LLMs to evaluate their own outputs has recently gained significant attention. Zheng et al. (2023) compared LLMs’ judgments with human evaluations on tasks like multi-turn dialogues and the Chatbot Arena. Hu et al. (2024) used perturbations to explore various aspects of text quality using prompt-based methods, but found that such methods often do not consistently provide reliable scores. Similarly, Huang

MTQ-Eval llama				MTQ-Eval aya			
Language	F1 (Top 10)	Language	F1 (Bottom 10)	Language	F1 (Top 10)	Language	F1 (Bottom 10)
eng_Latn	0.95	mal_Mlym	0.39	spa_Latn	0.97	kac_Latn	0.33
nld_Latn	0.80	khm_Khmr	0.39	pol_Latn	0.97	sot_Latn	0.33
ron_Latn	0.80	kan_Knda	0.37	eng_Latn	0.97	sna_Latn	0.33
nob_Latn	0.80	bod_Tibt	0.37	fra_Latn	0.95	kin_Latn	0.33
pes_Arab	0.77	amh_Ethi	0.37	ell_Grek	0.95	sin_Sinh	0.33
por_Latn	0.77	asm_Beng	0.37	deu_Latn	0.95	lin_Latn	0.33
swe_Latn	0.76	tel_Telu	0.36	ces_Latn	0.92	lug_Latn	0.33
nya_Latn	0.75	ben_Beng	0.36	ron_Latn	0.92	nso_Latn	0.33
zsm_Latn	0.74	ory_Orya	0.33	ita_Latn	0.92	pbt_Arab	0.33
ita_Latn	0.74	tam_Taml	0.33	nld_Latn	0.92	zul_Latn	0.33

Table 8: Top and bottom performing languages on Belebele text quality evaluation dataset.

MTQ-Eval llama		MTQ-Eval aya	
Script	% Improvement	Script	% Improvement
Khmr	30	Hebr	105
Mlym	18	Hans	94
Tibt	16	Taml	94
Deva	15	Geor	85
Ethi	15	Grek	83
Thai	15	Hant	82
Cyrl	14	Telu	70
Hans	13	Arab	65
Telu	12	Mymr	64
Mymr	9	Gujr	58
Hebr	9	Armn	58
Lao	8	Cyrl	57
Knda	6	Guru	55
Arab	5	Deva	54
Latn	5	Beng	49
Guru	0	Thai	48
Gujr	-2	Khmr	39
Sinh	-2	Mlym	39
Hant	-6	Lao	33
Grek	-6	Tibt	30
Geor	-11	Knda	21
Armn	-14	Ethi	21
Orya	-21	Latn	20
Taml	-21	Orya	0
Beng	-26	Sinh	0

Table 9: Improvement in F1 score (%) by MTQ-Eval over base model for scripts in Belebele dataset.

Lang.	llama		aya	
	SFT	MTQ-Eval	SFT	MTQ-Eval
ara	0.55	<b>0.71</b>	<b>0.67</b>	<b>0.67</b>
bul	<b>0.23</b>	0.13	0.19	<b>0.32</b>
bos	0.44	<b>0.62</b>	0.48	<b>0.62</b>
cze	<b>0.54</b>	0.28	<b>0.62</b>	0.47
deu	<b>0.44</b>	<b>0.44</b>	<b>0.41</b>	0.29
ell	0.19	<b>0.37</b>	<b>0.35</b>	0.21
eng	<b>0.80</b>	0.75	<b>0.80</b>	0.75
spa	0.58	<b>0.69</b>	0.31	<b>0.52</b>
fas	0.47	<b>0.55</b>	0.64	<b>0.75</b>
fra	<b>0.71</b>	0.49	<b>0.68</b>	0.64
heb	<b>0.40</b>	<b>0.40</b>	<b>0.47</b>	0.45
hin	0.42	<b>0.46</b>	<b>0.61</b>	0.42
hrv	0.38	<b>0.46</b>	0.16	<b>0.46</b>
hun	0.38	<b>0.5</b>	<b>0.48</b>	0.42
ita	0.43	<b>0.47</b>	0.24	<b>0.45</b>
jpn	0.52	<b>0.75</b>	<b>0.71</b>	0.61
lav	<b>0.43</b>	0.39	0.53	<b>0.54</b>
pol	<b>0.56</b>	<b>0.56</b>	0.77	<b>0.80</b>
por	<b>0.52</b>	0.49	<b>0.60</b>	<b>0.60</b>
rus	0.52	<b>0.6</b>	<b>0.53</b>	0.44
slk	0.55	<b>0.56</b>	0.28	<b>0.57</b>
slv	0.40	<b>0.45</b>	0.29	<b>0.49</b>
sqi	<b>0.26</b>	0.23	<b>0.40</b>	0.23
srp	<b>0.56</b>	0.54	0.19	<b>0.50</b>
swe	0.40	<b>0.51</b>	0.40	<b>0.41</b>
tha	0.58	<b>0.68</b>	0.48	<b>0.58</b>
urd	<b>0.57</b>	0.53	0.48	<b>0.71</b>
zho	<b>0.75</b>	0.64	0.66	<b>0.74</b>
Avg	0.49	<b>0.51</b>	0.48	<b>0.52</b>

Table 10: F1 scores on the MMS dataset.

et al. (2024) highlighted the limitations of fine-tuned judge models for LLM evaluation, noting that these models are typically task-specific and lack generalizability, while others questioned LLMs’ potential to replace human evaluators (Chiang and Lee, 2023).

Several prompt-based approaches have also been proposed although the findings have been mixed. For example, Hu et al. (2024) showed that varying the details in the evaluation criteria had little effect on LLM behavior. Murugadoss et al. (2024) found that detailed instructions are not always helpful in improving automatic evaluations made by LLMs. Doostmohammadi et al. (2024) noticed that prompt

based method works best mostly under few shot setting. In contrast, Hada et al. (2024a); Lee et al. (2024) demonstrated that detailed instructions resulted in evaluations that closely resembled human judgments. Chollampatt et al. (2025) leveraged machine translation and a reference LLM to generate a reference for comparison in high resource language. To our knowledge, we are the first to explore the use of DPO for aligning models with text quality preferences.

**Multilingual Evaluation** Recently, Sheng et al.



Model	Coh.	Con.	Flu.	Avg.
11ama SFT (LR)	3.25 $\pm$ 0.04	3.10 $\pm$ 0.11	2.99 $\pm$ 0.01	3.11
11ama MTQ-Eval (LR)	3.25 $\pm$ 0.04	3.18 $\pm$ 0.10	2.99 $\pm$ 0.01	<b>3.14</b>
11ama SFT (HR)	3.22 $\pm$ 0.04	3.18 $\pm$ 0.11	2.98 $\pm$ 0.01	<b>3.13</b>
11ama MTQ-Eval (HR)	3.20 $\pm$ 0.04	3.17 $\pm$ 0.11	2.98 $\pm$ 0.01	3.12
aya SFT (LR)	3.24 $\pm$ 0.04	3.09 $\pm$ 0.12	2.97 $\pm$ 0.02	3.10
aya MTQ-Eval (LR)	3.22 $\pm$ 0.04	3.13 $\pm$ 0.12	2.98 $\pm$ 0.01	<b>3.11</b>
aya SFT (HR)	3.22 $\pm$ 0.04	3.07 $\pm$ 0.11	2.98 $\pm$ 0.01	<b>3.09</b>
aya MTQ-Eval (HR)	3.24 $\pm$ 0.04	3.01 $\pm$ 0.11	2.98 $\pm$ 0.01	3.08

Table 11: Mean scores with 95% confidence intervals ( $\pm$  CI) for XLSum summaries. Confidence intervals were computed using the t-distribution and remain small and consistent across models.

Language	SFT Output	MTQ-Eval Output
English	A new study published in the journal Science finds that floods in some European regions, such as southern England, are occurring 15 days earlier than they did half a century ago due to climate change, while floods in other areas, such as the North Sea, are being delayed by around eight days. The study suggests that the changes in flood timing have significant implications for river ecosystems, hydropower stations, and agriculture, and may also impact the scale of flooding in the future.	A new study has found that floods in some European regions, such as southern England, are occurring 15 days earlier than they did 50 years ago due to climate change, while floods in other areas, like the North Sea, are being delayed by around 8 days. The changes in flood timing have significant implications for river ecosystems, hydropower stations, and agriculture.
Nepali	डाशङ्गि विमानस्थल ९८ वटा फुटबल मैदानजतरो ठूलो छ र चीनका राष्ट्रपति सी जिनपिङले बुधवार औपचारिक उद्घाटन गरे। यो विमानस्थल सञ्चालनमा आएपछि बेइजिङमा दुईवटा ठूला अन्तर्राष्ट्रिय विमानस्थल भएका छन्। Advertisement Advertisement ( <i>Daxing Airport, the size of 98 football fields, was officially opened by Chinese President Xi Jinping on Wednesday. With its opening, Beijing now has two major international airports. Advertisement Advertisement</i> )	चीनका राष्ट्रपति सी जिनपिङले बेइजिङस्थित डाशङ्गि अन्तर्राष्ट्रिय विमानस्थलको औपचारिक उद्घाटन गरेका छन्, जो ७ लाख वर्गमिटर क्षेत्रमा फैलिएको छ र ९८ वटा फुटबल मैदानजतरो ठूलो छ। यस विमानस्थलमा बुधवारदेखि सातवटा आन्तरिक विमानसेवाले उडान सुरु गरेका छन्। साथै अन्तर्राष्ट्रिय विमानसेवाहरूले पनि विमान सञ्चालन गर्ने घोषणा गरिसकेका छन्। ( <i>Chinese President Xi Jinping has officially opened Beijing's Daxing International Airport, which covers an area of 700,000 square meters and is as large as 98 football fields. Seven domestic airlines have started operating at the airport since Wednesday. International airlines have also announced their plans to operate flights.</i> )

Table 12: Summaries in English and Nepali generated by SFT and MTQ-Eval models for a news article.

(2024) proposed an evaluation metric requiring LLMs to exhibit strong proficiency in each language being assessed, while Trokhymovych et al. (2024) measured readability scores using a ranking-based fine-tuning approach. Hada et al. (2024b) evaluated multiple dimensions across 10 languages. In our study, we aim to enhance the capability of models to effectively differentiate between high- and low-quality text in a multilingual setting, expanding language coverage to 115 languages.

## 7 Conclusion

We present a new framework to enhance open-source LLMs’ capability to differentiate between high and low-quality text across multiple languages without the need for human annotations. Our approach demonstrates significant effectiveness across several datasets in various languages and also boosts downstream performance, that helps

models to better grasp the distinction between good and poor-quality text. Measurable improvements on downstream tasks suggest that our approach does not just benefit evaluation but also enhances performance in other tasks, opening exciting possibilities.

## Limitations

This paper does not cover all languages, leaving several unexplored. While our text degradation methods are effective and scalable for simulating certain types of errors (e.g., grammatical disruptions, loss of coherence), they do not capture other dimensions of low-quality text, such as semantic errors, factual inaccuracies, or nuanced linguistic issues. Furthermore, applying the same degradation methods uniformly across languages may not account for language-specific features. For instance, some languages (e.g., agglutinative or highly inflected languages) may not exhibit noticeable degra-

dation from simple word shuffling, and future research should exploit the linguistic diversity of the dataset. Additionally, resource constraints prevented us from experimenting with larger models.

## Ethics Statement

We could not examine whether the datasets used in this study contain biases toward certain languages, potentially introduced by human annotators.

## Acknowledgement

We thank Dr. Suresh Singh, PortNLP Lab members, and the anonymous reviewers for their invaluable feedback.

## References

- Ameeta Agrawal, Andy Dang, Sina Bagheri Nezhad, Rhitabrat Pokharel, and Russell Scheinberg. 2024. [Evaluating multilingual long-context models for retrieval and reasoning](#). In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 216–231, Miami, Florida, USA. Association for Computational Linguistics.
- Lukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, and Tomasz Jan Kajdanowicz. 2023. [Massively multilingual corpus of sentiment datasets and multifaceted sentiment classification benchmark](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Sina Bagheri Nezhad and Ameeta Agrawal. 2024. [What drives performance in multilingual language models?](#) In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 16–27, Mexico City, Mexico. Association for Computational Linguistics.
- Sina Bagheri Nezhad, Ameeta Agrawal, and Rhitabrat Pokharel. 2025. [Beyond data quantity: Key factors driving performance in multilingual language models](#). In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 225–239, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabisa. 2024. [The belebele benchmark: a parallel reading comprehension dataset in 122 language variants](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.
- Cheng-Han Chiang and Hung-yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.
- Shamil Chollampatt, Minh Quang Pham, Sathish Reddy Indurthi, and Marco Turchi. 2025. [Cross-lingual evaluation of multilingual text generation](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7766–7777, Abu Dhabi, UAE. Association for Computational Linguistics.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1).
- Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Ehsan Doostmohammadi, Oskar Holmström, and Marco Kuhlmann. 2024. [How reliable are automatic evaluation methods for instruction-tuned LLMs?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6321–6336, Miami, Florida, USA. Association for Computational Linguistics.
- Amila Ferron, Amber Shore, Ekata Mitra, and Ameeta Agrawal. 2023. [MEEP: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2078–2100, Singapore. Association for Computational Linguistics.
- Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Pudupully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023.

- Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Rishav Hada, Varun Gumma, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024a. **METAL: Towards multilingual meta-evaluation**. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2280–2298, Mexico City, Mexico. Association for Computational Linguistics.
- Rishav Hada, Varun Gumma, Adrian Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2024b. **Are large language model-based evaluators the solution to scaling up multilingual evaluation?** In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1051–1070, St. Julian’s, Malta. Association for Computational Linguistics.
- Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. **XLsum: Large-scale multilingual abstractive summarization for 44 languages**. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. **LoRA: Low-rank adaptation of large language models**. In *International Conference on Learning Representations*.
- Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. 2024. **Are LLM-based evaluators confusing NLG quality criteria?** In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9530–9570, Bangkok, Thailand. Association for Computational Linguistics.
- Hui Huang, Yingqi Qu, Jing Liu, Muyun Yang, and Tiejun Zhao. 2024. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers. *arXiv preprint arXiv:2403.02839*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.
- Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. **Mission: Impossible language models**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023. **Large language models are state-of-the-art evaluators of translation quality**. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.
- Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. **Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.
- Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, and Pilsung Kang. 2024. Checkeval: Robust evaluation framework using large language model via checklist. *arXiv preprint arXiv:2403.18771*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. **G-eval: NLG evaluation using gpt-4 with better human alignment**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. 2024. Evaluating the evaluator: Measuring llms’ adherence to task evaluation instructions. *arXiv preprint arXiv:2408.08781*.
- Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an llm to help with code understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.
- Rhitabrata Pokharel, Yufei Tao, and Ameeta Agrawal. 2025. Capo: Confidence aware preference optimization learning for multilingual preferences. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2025 (Findings)*. Association for Computational Linguistics.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. **Direct preference optimization: Your language model is secretly a reward model**. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Shuqian Sheng, Yi Xu, Tianhang Zhang, Zanwei Shen, Luoyi Fu, Jiaxin Ding, Lei Zhou, Xinbing Wang, and Chenghu Zhou. 2024. Repeval: Effective text evaluation with llm representation. *arXiv preprint arXiv:2404.19563*.

- Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. [Exploring LLM prompting strategies for joint essay scoring and feedback generation](#). In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.
- Mykola Trokhymovych, Indira Sen, and Martin Gerlach. 2024. [An open multilingual system for scoring readability of Wikipedia](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6296–6311, Bangkok, Thailand. Association for Computational Linguistics.
- Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Cl  mentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M Rush, and Thomas Wolf. 2024. [Zephyr: Direct distillation of LM alignment](#). In *First Conference on Language Modeling*.
- Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238.
- Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. [Is ChatGPT a good NLG evaluator? a preliminary study](#). In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore. Association for Computational Linguistics.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. [Neural network acceptability judgments](#). *Transactions of the Association for Computational Linguistics*, 7:625–641.
- Zhaofeng Wu, Ananth Balashankar, Yoon Kim, Jacob Eisenstein, and Ahmad Beirami. 2024. [Reuse your rewards: Reward model transfer for zero-shot cross-lingual alignment](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1332–1353, Miami, Florida, USA. Association for Computational Linguistics.
- Ziyin Zhang, Yikang Liu, Weifang Huang, Junyu Mao, Rui Wang, and Hai Hu. 2024. [MELA: Multilingual evaluation of linguistic acceptability](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2658–2674, Bangkok, Thailand. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

## A Appendix

### A.1 Dataset format for DPO

Figures 5 through 7 outline the prompts used in the dataset creation.

### A.2 Prompts used in different scenarios

Figures 8 through 10 outline the various prompts used in this study.

### A.3 Full results of top and bottom performing languages

Tables 13 presents the full results of top and bottom performing languages on Belebele text quality evaluation dataset.



[ 'Please act as an impartial judge and evaluate the text quality of the provided passage.\n\nText quality of a passage is defined by how well it maintains the following aspects.\n(1) Coherence - logical flow and connectivity between sentences and ideas in the text.\n(2) Fluency - smoothness and naturalness of individual sentences.\n(3) Simplicity - how easy it is to understand the passage.\n(4) Linguistic Acceptability - if the text sounds natural and correct to a native speaker.\n\nProvide a binary rating, "0" for low quality or "1" for high quality, strictly following this format: "[[0]]" or "[[1]]". Do not provide an explanation.\n\nPassage:\n\nL\'origami Pureland est un origami avec la contrainte qu\'un seul pli peut être fait à la fois, les plis plus complexes comme les plis inversés ne sont pas autorisés, et tous les plis ont des emplacements simples. Il a été développé par John Smith dans les années 1970 pour aider les personnes inexpérimentées ou ayant des capacités motrices limitées. ']

Figure 5: An example of *prompt* part of the DPO finetuning dataset

MTQ-Eval llama				MTQ-Eval aya			
Language	F1 (Top 20)	Language	F1 (Bottom 20)	Language	F1 (Top 20)	Language	F1 (Bottom 20)
eng_Latn	0.95	sin_Sinh	0.44	spa_Latn	0.97	tso_Latn	0.33
nld_Latn	0.80	guj_Gujr	0.44	pol_Latn	0.97	tsn_Latn	0.33
ron_Latn	0.80	npi_Deva	0.42	eng_Latn	0.97	yor_Latn	0.33
nob_Latn	0.80	hau_Latn	0.42	fra_Latn	0.95	xho_Latn	0.33
pes_Arab	0.77	kac_Latn	0.42	ell_Grek	0.95	ory_Orya	0.33
por_Latn	0.77	tir_Ethi	0.40	deu_Latn	0.95	grn_Latn	0.33
swe_Latn	0.76	fuv_Latn	0.40	ces_Latn	0.92	hat_Latn	0.33
nya_Latn	0.75	kat_Geor	0.39	ron_Latn	0.92	hau_Latn	0.33
zsm_Latn	0.74	lao_Lao	0.39	ita_Latn	0.92	ibo_Latn	0.33
ita_Latn	0.74	urd_Arab	0.39	nld_Latn	0.92	ssw_Latn	0.33
fra_Latn	0.73	mal_Mlym	0.39	ind_Latn	0.9	kac_Latn	0.33
spa_Latn	0.73	khm_Khmr	0.39	por_Latn	0.9	sot_Latn	0.33
isl_Latn	0.72	kan_Knda	0.37	ukr_Cyrl	0.87	sna_Latn	0.33
hat_Latn	0.72	bod_Tibt	0.37	pes_Arab	0.87	kin_Latn	0.33
ilo_Latn	0.72	amh_Ethi	0.37	vie_Latn	0.87	sin_Sinh	0.33
als_Latn	0.70	asm_Beng	0.37	arb_Arab	0.87	lin_Latn	0.33
uzn_Latn	0.70	tel_Telu	0.36	cat_Latn	0.85	lug_Latn	0.33
jav_Latn	0.70	ben_Beng	0.36	slk_Latn	0.82	nso_Latn	0.33
slk_Latn	0.69	ory_Orya	0.33	lit_Latn	0.80	pbt_Arab	0.33
azj_Latn	0.69	tam_Taml	0.33	heb_Hebr	0.80	zul_Latn	0.33

Table 13: Top and bottom (20) performing languages on Belebele text quality evaluation dataset.

```

[[{'content': 'Please act as an impartial judge and evaluate the text
quality of the provided passage.\n\nText quality of a passage is defined
by how well it maintains the following aspects.\n(1) Coherence - logical
flow and connectivity between sentences and ideas in the text.\n(2)
Fluency - smoothness and naturalness of individual sentences.\n(3)
Simplicity - how easy it is to understand the passage.\n(4) Linguistic
Acceptability - if the text sounds natural and correct to a native
speaker.\n\nProvide a binary rating, "0" for low quality or "1" for high
quality, strictly following this format: "[[0]]" or "[[1]]". Do not
provide an explanation.\n\nPassage:\n\nL\'origami Pureland est un origami
avec la contrainte qu\'un seul pli peut être fait à la fois, les plis
plus complexes comme les plis inversés ne sont pas autorisés, et tous
les plis ont des emplacements simples. Il a été développé par John
Smith dans les années 1970 pour aider les personnes inexpérimentées ou
ayant des capacités motrices limitées.',
'role': 'user'},
{'content': 'Please act as an impartial judge and evaluate the text
quality of the provided passage.\n\nText quality of a passage is defined
by how well it maintains the following aspects.\n(1) Coherence - logical
flow and connectivity between sentences and ideas in the text.\n(2)
Fluency - smoothness and naturalness of individual sentences.\n(3)
Simplicity - how easy it is to understand the passage.\n(4) Linguistic
Acceptability - if the text sounds natural and correct to a native
speaker.\n\nProvide a binary rating, "0" for low quality or "1" for
high quality, strictly following this format: "[[0]]" or "[[1]]". Do
not provide an explanation.\n\nPassage:\n\nL\'origami Pureland est un
origami avec la contrainte qu\'un seul pli peut être fait à la fois, les
plis plus complexes comme les plis inversés ne sont pas autorisés, et tous
les plis ont des emplacements simples. Il a été développé par John Smith
dans les années 1970 pour aider les personnes inexpérimentées ou ayant
des capacités motrices limitées. \nRating: [[1]]',
'role': 'assistant'}]]

```

Figure 6: An example of *chosen* part of the DPO finetuning dataset

```

[[{'content': 'Please act as an impartial judge and evaluate the text
quality of the provided passage.\n\nText quality of a passage is defined
by how well it maintains the following aspects.\n(1) Coherence - logical
flow and connectivity between sentences and ideas in the text.\n(2)
Fluency - smoothness and naturalness of individual sentences.\n(3)
Simplicity - how easy it is to understand the passage.\n(4) Linguistic
Acceptability - if the text sounds natural and correct to a native
speaker.\n\nProvide a binary rating, "0" for low quality or "1" for high
quality, strictly following this format: "[[0]]" or "[[1]]". Do not
provide an explanation.\n\nPassage:\n\nL\'origami Pureland est un origami
avec la contrainte qu\'un seul pli peut être fait à la fois, les plis
plus complexes comme les plis inversés ne sont pas autorisés, et tous les
plis ont des emplacements simples. Il a été développé par John Smith
dans les années 1970 pour aider les personnes inexpérimentées ou ayant
des capacités motrices limitées.',
  'role': 'user'},
{'content': 'Please act as an impartial judge and evaluate the text
quality of the provided passage.\n\nText quality of a passage is defined
by how well it maintains the following aspects.\n(1) Coherence - logical
flow and connectivity between sentences and ideas in the text.\n(2)
Fluency - smoothness and naturalness of individual sentences.\n(3)
Simplicity - how easy it is to understand the passage.\n(4) Linguistic
Acceptability - if the text sounds natural and correct to a native
speaker.\n\nProvide a binary rating, "0" for low quality or "1" for high
quality, strictly following this format: "[[0]]" or "[[1]]". Do not
provide an explanation.\n\nPassage:\n\nL\'origami Pureland est un origami
avec la contrainte qu\'un seul pli peut être fait à la fois, les plis
plus complexes comme les plis inversés ne sont pas autorisés, et tous les
plis ont des emplacements simples. Il a été développé par John Smith
dans les années 1970 pour aider les personnes inexpérimentées ou ayant
des capacités motrices limitées. \nRating: [[0]]',
  'role': 'assistant'}]]

```

Figure 7: An example of *rejected* part of the DPO finetuning dataset

Determine whether the following sentence(s) violate certain linguistic constraints. If yes, then it is "unacceptable"; otherwise, "acceptable".

Sentence: {sentence}

Determine whether this sentence is acceptable or unacceptable? If acceptable, return [[1]], otherwise [[0]]. Strictly follow this format: "[[0]]" or "[[1]]". Do not provide any feedback.

Your Answer:

Figure 8: Prompt used for MELA evaluation.

Please act as an impartial judge and determine the sentiment of the following passage.

Passage: {text}

If the sentiment is negative, return `[[0]]`, if neutral, return `[[1]]`, and if positive, return `[[2]]`. Strictly follow this format: `"[[0]]"`, `"[[1]]"`, or `"[[2]]"`. Do not provide any feedback.

Your Answer:

Figure 9: Prompt used for MMS evaluation.

Summarize the following article in one or two sentence.

Article:  
{text}

Do not include any additional note or text; simply output one sentence summary and nothing more.

Your Answer:

Figure 10: Prompt used for XLSum summary generation.