XQ-MEval: A Dataset with Cross-lingual Parallel Quality for Benchmarking Translation Metrics

Anonymous ACL submission

Abstract

Automatic evaluation metrics are essential for building multilingual translation systems. The common practice of evaluating these systems is averaging metric scores across languages, yet this is suspicious since metrics may suffer from cross-lingual scoring bias, where translations of equal quality receive different scores across languages. This problem has not been systematically studied because no benchmark exists that provides parallel-quality instances across languages, and expert annotation is not realistic. In this work, we propose XQ-MEval, a semi-automatically built dataset covering nine translation directions, to benchmark translation metrics. Specifically, we inject MQM-defined errors into gold translations automatically, filter them by native speakers for reliability, and merge errors to generate pseudo translations with controllable quality. These pseudo translations are then paired with corresponding sources and references to form triplets used in assessing the qualities of translation metrics. Using XQ-MEval, our experiments on nine representative metrics reveal the inconsistency between averaging and human judgment and provide the first empirical evidence of cross-lingual scoring bias. Finally, we propose a normalization strategy derived from XQ-MEval that aligns score distributions across languages, improving the fairness and reliability of multilingual metric evaluation.[1]

1 Introduction

With the growing demand for multilingual translation systems, comprehensive and reliable evaluation has become critical (Kocmi et al., 2024). In human evaluation, Multidimensional Quality Metrics (MQM) largely achieves cross-lingually comparable evaluation through standardized error categories and hierarchical deduction (Lommel et al., 2013; Freitag et al., 2021). However, as evaluation scales up, automatic evaluation metrics are essential due to their efficiency and scalability (Popović, 2015, 2017; Post, 2018; Goyal et al., 2022). Therefore, MQM driven automatic metrics have recently become the primary tools, e.g., COMET (Rei et al., 2020) and MetricX (Juraska et al., 2023).

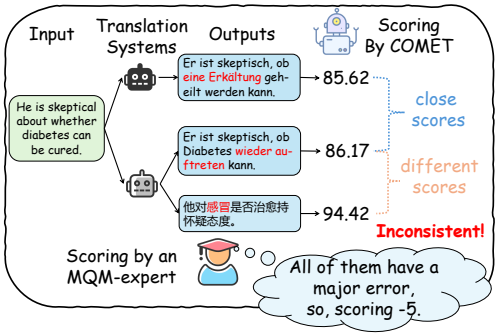

Figure 1: A clue of this study, showing the inconsistency between human evaluation, i.e., MQM, and automatic metrics, e.g., COMET. Three translations each contain one major error, thus sharing the same MQM score, yet COMET assigns notably different scores, with larger gaps across languages.

In multilingual translation evaluation, the common practice is to evaluate each language direction with a metric and then average the metric scores to compute a system-level score[2] (Chen et al., 2023; Cao et al., 2024; Qu et al., 2025). However, this average strategy may be problematic because it implicitly assumes that different languages are scored on the same scale for a similar error. In fact, cross-lingual scoring bias is indeed observed as illustrated in Figure 1. To quantify and verify this potential problem, a benchmark is needed that provides parallel quality across languages, ensuring that cross-lingual comparisons are made on the same grounds, i.e., similar errors are quantified equally across different languages. Due to the unaffordable cost of expert-level annotations, no such benchmark currently exists.

In this work, we propose a novel semi-automatic

[1] We will release our codes and dataset after acceptance.

[2] The computational procedure of the average strategy is described in Appendix A with pseudocode.

pipeline that injects MQM-defined errors into gold translations and filters them with native speakers, ensuring reliability and cross-lingual consistency. By merging individual errors, we generate pseudo translations with controllable quality, which are then paired with gold sources and references to form triplets. Based on this, we construct a dataset for evaluating metrics with cross-lingual parallel quality, namely XQ-MEval. This dataset covers nine languages[3], i.e., Chinese, Japanese, Lao, Vietnamese, Indonesian, French, Spanish, Sinhala, and German, for translation directions from English, and provides parallel-quality triplets for the fair metric comparisons across languages.

Based on XQ-MEval, we conduct experiments on nine representative automatic metrics. The results reveal a clear inconsistency between averaging and human evaluation, and provide the first empirical evidence of cross-lingual scoring bias. This bias has two manifestations: (1) systems of equal quality receive different scores across languages; (2) the decline of metric scores with decreasing quality is inconsistent across languages. Building on this finding, we propose a simple strategy based on normalization (García et al., 2015), i.e., Language-specific Global Normalization (LGN), to calibrate multilingual evaluation metrics. Our experiments show that, compared to the average strategy, LGN effectively reduces score range disparities and improves the fairness and reliability of multilingual metric evaluation. We make the following threefold contributions in this study:

- We present XQ-MEval, the first multilingual dataset with parallel-quality triplets across nine translation directions, enabling benchmarking of automatic evaluation metrics.
- We evaluate representative metrics to reveal the inconsistency between the average strategy and human judgment, and provide the first analysis of cross-lingual scoring bias.
- We introduce and verify LGN, a normalized average strategy that calibrates metrics in evaluating multilingual translation systems.

2 Related Work

The evaluation of bilingual translation systems relies on discrete scoring schemes (Koehn and Monz, 2006; Vilar et al., 2007; Callison-Burch et al., 2007; Denkowski and Lavie, 2010), but these suffer from low inter-annotator agreement. Although Graham et al. (2013); Bojar et al. (2016, 2017) introduced the continuous rating scale to mitigate this variability, subjectivity-related biases persisted across annotators. Building upon the Multidimensional Quality Metrics (MQM) proposed by Lommel et al. (2013), Freitag et al. (2021) developed a framework that reduces annotator inconsistency through standardized error categories and hierarchical deduction. Specifically, each sentence is assumed to have perfect quality initially, and points are deducted according to error type, e.g., accuracy and fluency, and severity, e.g., 1 for minor and 5 for major. This makes translation metrics cross-lingually comparable because sentences with the same errors are expected to receive the same score across languages.

To complement costly and inconsistent human-based evaluation, automatic evaluation metrics are proposed to approximate human judgments of translation quality efficiently. They can be broadly categorized into three types: (1) *Regression-based metrics* frame evaluation as a supervised task that directly predicts scalar quality scores, including both models trained explicitly for evaluation, e.g., COMET (Rei et al., 2020, 2022a; Guerreiro et al., 2024) and MetricX (Juraska et al., 2023, 2024), and converting LLMs into evaluators, e.g., ReMedy (Tan and Monz, 2025). (2) *Sequence-based metrics* evaluate translations by comparing candidate translations with gold references, primarily relying on surface-level similarity[4], e.g., BLEU (Papineni et al., 2002; Post, 2018) and chrF (Popović, 2015, 2017). (3) *Reference-free metrics*, also known as quality estimation (QE), extend regression-based methods to evaluate translations directly against the source without requiring references, e.g., COMET-kiwi (Rei et al., 2021, 2023). In parallel, recent work has explored using LLMs as human evaluators by prompting them to follow explicit assessment agreements such as MQM, thereby approximating human judgment behavior at inference time (Kocmi and Federmann, 2023).

These metrics are widely applied in multilingual translation evaluation, but the practice of averaging scores across languages (Zhang et al., 2021; Qu and Watanabe, 2022; Chen et al., 2023; Cao et al., 2024; Qu et al., 2025) may hinder the system-

[3]Appendix B shows the details in language selection.

[4]Although metrics like BLEURT (Sellam et al., 2020) are regression-based, the metric depended on embeddings from sequence information should be classified as sequence-based.

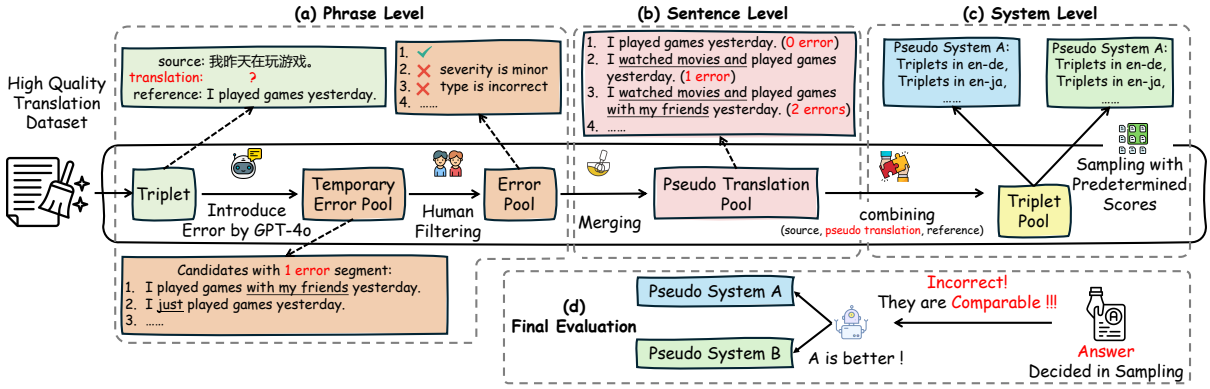

Figure 2: The illustration of our pipeline. Specifically, stages from (a) to (c) show the data construction and reveal that the product is to create pseudo translation systems with predetermined scores. Finally, stage (d) demonstrates the use of pseudo systems to assess the automatic metrics based on the answer, i.e., the predetermined score.

level evaluation since it is unclear whether a similar error is consistently measured across languages. Von Däniken et al. (2025) showed that metrics fail to align with human evaluation even in a single translation direction. Thus, benchmarks are needed to expose cross-lingual scoring bias and guide metric improvement. However, constructing them incurs costs similar to MQM, where each instance requires expert-level annotation. Fortunately, using LLMs with human filtering can simplify this process (Li et al., 2023; Kwan et al., 2024; Bai et al., 2024; Wang et al., 2025), providing a practical avenue for benchmark construction.

3 Pipeline of Dataset Construction

We present a multilingual dataset, XQ-MEval, for benchmarking automatic evaluation metrics covering nine translation directions, i.e., `en-zh`, `en-ja`, `en-lo`, `en-vi`, `en-id`, `en-fr`, `en-es`, `en-si`, and `en-de`, comprising both high-resource and low-resource languages[5]. Constructing such a dataset following MQM is challenging due to the high cost of expert annotation, which greatly limits the language coverage. To address this, we employ a semi-automatic approach, formatting each sample as a triplet and rigorously controlling quality to ensure cross-lingual parallelism. This design enables flexible sampling to simulate systems with predetermined quality levels for metric benchmarking.

Specifically, we introduce a novel pipeline for benchmark construction that enables systematic and cost-effective analysis of metric biases in Figure 2, comprising phrase-level, sentence-level, and system-level stages of different granularity. Automatic evaluation metrics operate on a triplet comprising a source, translation, and reference. We begin with a high-quality translation corpus, where each translation pair forms the source and reference for a triplet. At the phrase-level stage, a major-severity error is introduced into each reference. Then, at the sentence-level stage, we merge 0 to 5 errors from such candidates[6] to generate pseudo translations[7] with six distinct quality levels. Finally, at the system-level stage, pseudo systems are constructed by assembling triplets across different quality levels, thereby emulating translation systems with predetermined performance.

Nevertheless, we acknowledge that XQ-MEval instances are synthesized rather than produced by real translation systems, and may thus differ from real-world scenarios. We have conducted preliminary experiments on usable real-world MQM datasets and validated our approach in Appendix C.

3.1 Phrase-level Construction

XQ-MEval is built on Flores[8], a high-quality multilingual translation dataset, denoted as $\mathbb{F}$, with 102 instances used in our experiments[9]. Flores is particularly suitable because its translations are semantically parallel and are carefully validated by multiple native speakers (NLLB Team, 2022).

As shown in Figure 2, we define each translation instance in $\mathbb{F}$ as (s, r) where s represents the source in `en` and r represents its reference. We em-

[5]Languages are represented by ISO 639-1 codes, and details about language selection are shown in the Appendix B.

[6]The choice of 5 follows Google's MQM guideline, where each sentence can lose at most 25 points and each major error accounts for 5 points (Freitag et al., 2021).

[7]Annotators' feedback indicates that although combining errors may appear unnatural, they remain objectively valid.

[8]https://huggingface.co/datasets/openlanguagedata/flores_plus

[9]We have manually selected to exclude very short sentences that cannot accommodate multiple injected errors.

Part	Product	Operation	Language	Example	Note
1	Error Pool	Introducing single error to the reference by GPT-4o, then filter by native speakers.	de	Der Klage zufolge wurde der Abfall aus dem UN-Lager nicht ordnungsgemäß gesäubert, was dazu führte, dass Bakterien in den Zufluss des Artibonit-Flusses, einem der größten Flüsse Haitis, <v>sowie in andere Gewässer</v> gelangten.	Error type: Addition; Human judgment: ✔
			ja	訴訟によれば、国連キャンプの廃棄物が適切に消毒されていなかったため、ハイチ最大級の<v></v>に細菌が侵入したとのことです。	Error type: Omission; Human judgment: ✔
			zh	诉讼材料显示，联合国营地未能对废弃物进行<v> 彻底的焚烧 </v>，因而导致细菌进入阿蒂博尼特河的支流，这是海地最大河流之一。	Error type: Mistranslation; Human judgment: ✔
			zh	诉讼材料显示，联合国营地未能对废弃物进行适当的消毒处理，因而导致细菌进入<v>Artibonite River</v>的支流，这是海地最大河流之一。	Error type: Untranslated; Human judgment: ✘
2	Pseudo Translation Pool	Merging several candidates, where the error span is not conflict, to get a pseudo translation.	ja	訴訟によれば、国連キャンプの<v> 食料供給 </v>が適切に消毒されていなかったため、ハイチ最大級のアルティボナイト川の支流に細菌が侵入したとのことです。	Error span: 14-18
			ja	訴訟によれば、国連キャンプの廃棄物が適切に消毒されていなかったため、<v> さらに多くの問題が発生し、</v>ハイチ最大級のアルティボナイト川の支流に細菌が侵入したとのことです。	Error span: 34-35
			ja	訴訟によれば、国連キャンプの<v> 食料供給 </v>が適切に消毒されていなかったため、<v> さらに多くの問題が発生し、</v>ハイチ最大級的阿蒂博尼特川的支流に細菌が侵入したとのことです。	Pseudo Translation with 2 Errors
3	Triplet Pool	Each triplet fed into metrics is combined by a source, a translation, and a reference.	en	According to the lawsuit, waste from the UN camp was not properly sanitized, causing bacteria to enter the tributary of the Artibonite River, one of Haiti's largest.	Source
			ja	訴訟によれば、国連キャンプの<v> 食料供給 </v>が適切に消毒されていなかったため、<v> さらに多くの問題が発生し、</v>ハイチ最大級的阿蒂博尼特川的支流に細菌が侵入したとのことです。	Pseudo Translation
			ja	訴訟によれば、国連キャンプの廃棄物が適切に消毒されていなかったため、ハイチ最大級的阿蒂博尼特川的支流に細菌が侵入したとのことです。	Reference

Table 1: Examples used to assist in explaining Figure 2. The column of part is used for conveniently referring.

ploy GPT-4o[10] (OpenAI, 2024) to inject an MQM-defined error of major severity into r, producing a temporary error candidate $\hat{r}$ comprising a single error segment with an identification tag.

We introduce the following four error types, which dominate existing MQM datasets[11] and are conducive to cross-lingual comparability as they are purely semantic (Haspelmath, 2010; Cristofaro, 2009): (1) *Addition*, where extraneous information is inserted in translations; (2) *Omission*, where a part of the source is left out; (3) *Mistranslation*, where the meaning is distorted or incorrect; (4) *Untranslated*, where the source remains untranslated text. Because each pseudo translation $\tilde{r}$ may contain up to five errors in our settings, we allow multiple instances of the same error type injected separately into the first and second halves of the sentence, which are first divided and explicitly tagged to guide GPT-4o to introduce error segments into the corresponding parts. Thus, a single (s, r) can yield up to eight temporary error candidates $\hat{\boldsymbol{r}} = \{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_8\}$. Applying this process to the entire dataset produces a temporary error pool $\hat{\mathbb{R}} = \bigcup_{i=1}^{n} \hat{\boldsymbol{r}}_i$.[12]

[10]Version: *gpt-4o-2024-11-20*.

[11]These four types account for 46.3% of all MQM errors.

[12]Prompts are carefully designed and listed in Appendix E.

	en-zh	en-lo	en-ja	en-vi	en-id	en-fr	en-es	en-si	en-de
Add.	194	196	194	201	196	200	196	200	203
Omit.	200	191	196	199	197	196	197	188	194
Mist.	201	200	202	200	201	196	197	197	194
Untr.	181	172	183	171	188	183	181	181	183

Table 2: The number of candidates generated by GPT-4o and filtered by annotators for each error type. The abbreviations of error type are as follows: Addition, Omission, Mistranslation, and Untranslated.

Then, native speakers of the nine target languages review and filter $\hat{\mathbb{R}}$. In practice, two independent reviewers are engaged, but, for `si`, `lo`, and `vi`, only one reviewer is available due to resource constraints. Finally, only $\hat{r}$ unanimously approved by both annotators are retained to construct the final error pool $\hat{\mathbb{R}}_{\text{filtered}}$. The part 1 of Table 1 demonstrates this process.

To ensure consistency, we provide detailed annotation guidelines in Appendix D that explain the four MQM errors and specify filtering conditions regarding completeness, locality, and severity. Table 2 summarizes the number of sentences generated by GPT-4o and retained by annotators for each error type. Also, to assess annotation reliability, we compute inter-annotator agreement between the two native speakers. As shown in Table 3, agreement is consistently high, reflecting the effectiveness of our guidelines. We further validate robustness through a second round of independent

	en-zh	en-ja	en-fr	en-es	en-de	en-id
Agreement (%)	98.16	96.45	97.79	97.30	97.67	96.45

Table 3: The annotation agreement between the two native speakers during the manual screening process.

	en-zh	en-lo	en-ja	en-vi	en-id	en-fr	en-es	en-si	en-de
max	176	176	150	218	176	139	139	139	176
min	19	8	11	21	7	8	11	11	15

Table 4: Summarizes the maximum and minimum number of pseudo translations generated for each triplet in different translation directions.

screening on 200 randomly sampled en-zh and en-ja instances. The alignment rates between the two rounds are 99% for en-zh and 98% for en-ja, confirming the stability of annotation process. These results demonstrate that the constructed dataset is both reliable and reproducible, establishing a solid foundation for subsequent stages.

3.2 Sentence-level Construction

Based on $\hat{\mathbb{R}}_{\text{filtered}}$, we generate each pseudo translation $\tilde{r}$ by merging k single-error candidates $\hat{r}$, where $k \in \{0, 1, 2, 3, 4, 5\}$, all of which are from the same $\hat{\boldsymbol{r}}_{\text{filtered}}$, i.e., the candidates filtered for each pair (s, r), as illustrated in Figure 2. $\tilde{r}$ is a variant of r containing between 0 and 5 errors, thus covering six distinct quality levels in the MQM framework. Part 2 of Table 1 provides an example, where two non-overlapping $\hat{r}$ are merged to form a $\tilde{r}$ with two errors. In addition, a special case is that of 0 error, corresponding to the reference itself.[13]

By merging candidates, we can flexibly produce pseudo translations with the desired scores. However, candidates may contain overlapping error spans, which compromise the locality of each error. Such overlapping combinations are simply discarded so that the actual number of pseudo translations is smaller than the theoretical maximum. As a result, each triplet yields a set of pseudo translations that cover different quality levels. Table 4 reports the minimum and maximum number of pseudo translations generated per triplet for each language direction, reflecting the constraints imposed by overlap and sentence structure.

3.3 System-level Construction and Final Evaluation

As shown in part 3 of Table 1, an instance is formed as a triplet $(s, \tilde{r}, r)$. By iterating over the entire dataset, we obtain the triplet pool $\mathcal{D}$, which constitutes the final dataset of XQ-MEval.

[13]In this case, a metric should assign a full score to the triplet, when the translation matches the gold reference.

Figure 2 further illustrates how $\mathcal{D}$ enables systematic benchmarking of automatic metrics. We assume the existence of a translation system with a given MQM score derived from the number of error spans and then construct a pseudo system by sampling triplets that reflect this target performance. This procedure is both flexible and powerful because it allows us to generate arbitrary pseudo systems tailored to different evaluation scenarios. Based on pseudo systems with predefined performance, we evaluate them using automatic metrics and measure the alignment between metric scores and predefined scores as a proxy for consistency with human judgments.[14]

4 Experimental Setup

Based on XQ-MEval in Section 3, we perform a large-scale and multilingual analysis of existing automatic evaluation metrics[15] as follows.

Sequence-based (1) spBLEU (Goyal et al., 2022), a variant of BLEU that unifies tokenization across languages through a SentencePiece tokenizer (Kudo and Richardson, 2018); (2) chrF++ (Popović, 2017), which assesses character-level overlap and balances precision with recall; (3) BLEURT-20 (Sellam et al., 2020), a BERT-based metric trained on human-annotated data to better align with human judgments.

Regression-based (1) COMET-22 (Rei et al., 2022a), which integrates source, hypothesis, and reference embeddings to predict quality scores; (2) xCOMET-XL (Guerreiro et al., 2024), which improves interpretability by detecting errors explicitly; (3) MetricX-23 (Juraska et al., 2023), abbreviated as MX-reg, initialized with mT5 (Xue et al., 2021) and fine-tuned on MQM data.

Reference-free (1) COMET-KIWI-22 (Rei et al., 2022b), abbreviated as KIWI22, a reference-free variant of COMET-22; (2) COMET-KIWI-23 (Rei et al., 2023), abbreviated as KIWI23, an extended version of KIWI22; (3) MetricX-23-QE (Juraska et al., 2023), abbreviated as MX-qe, the reference-free variant of MetricX-23.

[14]Appendix A exhibits the process of computing system-level metric scores, and shows comparing them to predefined scores, i.e., human evaluations.

[15]We primarily focus on metrics within the categories defined in Section 2. However, we also analyze LLM-based approaches, including LLM-adapted regression metrics and MQM-style LLM-as-judge evaluation, in Appendix J.

Num. of Lang.	System-level Kendall-τ							Triplet-level Kendall-τ						
	BLR.	COM.	xCOM.	MX-r.	KW22.	KW23.	MX-q.	BLR.	COM.	xCOM.	MX-r.	KW22.	KW23.	MX-q.
3	0.89	0.88	0.90	0.80	0.88	0.86	0.82	0.50	0.46	0.38	0.35	0.44	0.39	0.32
6	0.88	0.88	0.89	0.84	0.90	0.89	0.83	0.46	0.44	0.42	0.38	0.45	0.39	0.33
9	0.87	0.89	0.90	0.83	0.90	0.89	0.83	0.48	0.42	0.44	0.38	0.44	0.38	0.32

Table 5: Results showing the system-level, and triplet-level Kendall-τ correlation between averaged metric scores and human judgments on pseudo systems. Num. of Lang. denotes the number of involved languages. In this setting, Num. of 3 means that the system is sampled from `zh`, `lo`, and `de`; Num. of 6 means that the system is sampled from `zh`, `lo`, `de`, `ja`, and `si`; Num. of 9 means that the system is sampled from all languages. The abbreviations of metric are as follows: BLEURT, COMET, xCOMET, MX-reg, KIWI22, KIWI23, and MX-qe.

Quality Level	spBLEU	chrF	BLEURT	COMET	xCOMET	MX-reg	KIWI22	KIWI23	MX-qe
1	1.53	7.56	1.62	2.51	9.95	22.80	2.38	6.05	23.61
2	3.16	9.84	4.46	3.84	14.77	23.74	2.39	8.55	22.38
3	5.04	12.00	7.26	5.09	20.51	23.75	2.66	11.23	20.57
4	7.25	14.22	10.07	6.27	26.01	24.23	3.29	14.42	18.94
5	10.01	16.23	13.79	7.61	28.67	24.03	3.76	18.86	15.40

Table 6: Illustration of the cross-lingual CV (%) of scores for nine automatic metrics measured at five quality levels.

5 Analysis on Average Strategy

5.1 Verification

To verify the consistency between the average strategy and human evaluations in multilingual MT evaluation, we assemble 10 pseudo systems to approximate real-world translation systems.

Following the procedure of Section 3.3, each pseudo system is built by aggregating 102 triplets sampled per language pair from multiple languages to meet predetermined scores. After scoring each triplet, system-level metric scores are computed by averaging their respective scores across directions, followed by calculating their correlation with human evaluation to assess agreement. This procedure is repeated 100 times for stability, and the average correlation across these repetitions is reported. We rely on the Kendall-τ coefficient (Kendall, 1938), a statistical measure of rank correlation, to quantify the consistency between the rankings induced by metrics and by predetermined scores, where higher values indicate stronger consistency and vice versa.

Table 5 reports the system-level correlation results under three settings with 3, 6, and all 9 languages, where the subsets of 3 and 6 were selected to maximize linguistic diversity. Although correlations appear high across settings, this is expected in our simplified evaluation setup, where instance quality is divided into five coarse-grained levels with large gaps, making quality differences easier for metrics to distinguish. As a result, such high correlations may be inflated by the evaluation setup and should be interpreted with caution.

To further examine whether this apparent consistency holds at a finer granularity, we analyze metric behavior at the triplet level. Since pseudo systems are constructed from triplets, we group all possible triplets across languages to form test systems. Table 5 presents the resulting triplet-level correlations, which are substantially lower and indicate pronounced inconsistency. These results shed light on the concerns raised by the system-level analysis and point to potential cross-lingual inconsistencies in metric scoring behavior.

5.2 Analysis

To analyze inconsistencies between metrics and human evaluations, we construct pseudo monolingual systems, each restricted to a single translation direction and quality level. Unlike multilingual systems, this setting isolates metric behavior within one language and enables direct cross-language comparison at the same quality level. Moreover, to address imbalances in triplet counts across quality levels[16], we randomly sample 102 triplets per system and repeat this procedure 10 times to ensure robustness.[17]

At the same quality level Table 6 reports cross-lingual coefficients of variation (CV) for nine metrics across five quality levels, corresponding to translations with the number of errors ranging from 1 to 5. For each quality level, CV is computed from the mean and standard deviation of metric scores across nine monolingual systems. CV measures score inconsistency across languages at the same quality level, indicating whether metrics provide consistent judgments as translation direction

[16]Appendix F counts and lists the triplets distribution of different languages.

[17]We further report tests with 5, 10, and 25 repetitions in Appendix G to support our design choices.

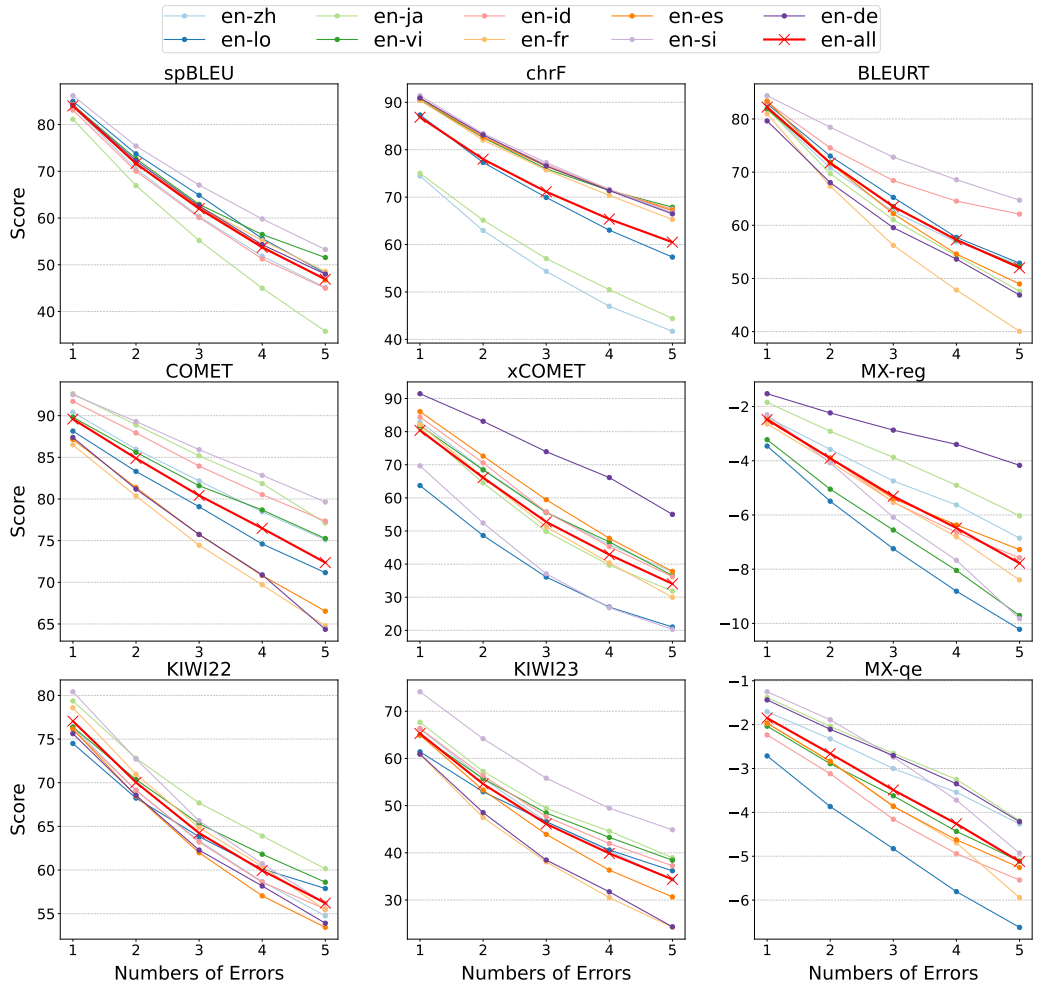

Figure 3: Visualization of nine metrics scores across nine directions at varying translation quality levels. `en-all` denoting the average metric scores among all directions.

varies, with ideal values close to zero. Results show inconsistencies for most metrics, with CV increasing as translation quality decreases. This indicates that metrics assign divergent scores to translations of comparable quality, deviating from human evaluation and reflecting cross-lingual bias in the scoring behavior of metrics.

Across different quality level Figure 3 plots metric scores across translation directions at varying quality levels to examine whether score trends remain consistent as quality varies.[18] Curves across directions should overlap, with similar scores and trends across quality levels. In contrast, two phenomena are observed.[19] First, metric scores differ across directions even at the same quality level. Second, as quality decreases, score reduction rates vary across directions, leading to widening gaps between curves. Consistent with the analysis in Table 6, these variations confirm the existence of cross-lingual scoring bias in automatic translation metrics, posing a challenge for metrics to align with human evaluations in multilingual settings, where uniformity across directions is expected.

6 Normalization-based Scoring

6.1 Methodology

The analysis in Section 5.2 reveals substantial variation in metric score ranges across translation directions. Figure 4 further illustrates this issue using COMET, where the distribution of scores for different target languages diverges even when the human score is fixed at 15, comprising 3 errors in each translation. It is evident that different languages occupy distinct numerical scales, making metric scores inconsistent even when human quality is comparable.

[18] Specific values are provided in Appendix H.

[19] Appendix I describes the difference across directions and across metrics in detail.

Num. of Lang.	System-level Kendall-τ							Triplet-level Kendall-τ						
	BLR.	COM.	xCOM.	MX-r.	KW22.	KW23.	MX-q.	BLR.	COM.	xCOM.	MX-r.	KW22.	KW23.	MX-q.
3	**0.90**	**0.89**	**0.91**	**0.88**	**0.90**	**0.88**	**0.85**	**0.51**	**0.48**	**0.47**	**0.41**	**0.45**	**0.41**	**0.34**
6	**0.92**	**0.91**	**0.90**	**0.91**	**0.92**	**0.90**	**0.86**	**0.49**	**0.48**	**0.48**	**0.42**	**0.46**	**0.41**	**0.35**
9	**0.91**	**0.93**	**0.92**	**0.88**	**0.91**	**0.91**	**0.86**	**0.50**	**0.48**	**0.49**	**0.41**	**0.45**	**0.40**	**0.34**

Table 7: Kendall-τ correlations at system-level and triplet-level, corresponding to Table 5. All settings and abbreviations follow Table 5. Bold values indicate improvements of LGN over the average strategy. Improvements are modest in magnitude but statistically significant; significance tests are reported in the Appendix K.

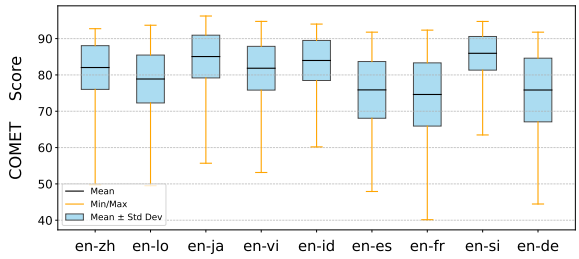

Figure 4: The illustration of COMET score distribution across different translation directions under fixed human evaluation scores. The bar sections represent the mean $\pm$ standard deviation, while the whiskers indicate the maximum and minimum values.

To address this problem, we propose **Language-specific Global Normalization** (LGN), which adopts z-score normalization to unify score scales across languages via mean and standard deviation. LGN computes the mean and standard deviation of triplet scores for each translation direction across all quality levels. For a given direction, 102 triplets are randomly sampled per quality level (including error-free translations) and pooled to calculate the global mean and standard deviation.[20] This process is repeated 10 times, and the final values are obtained by averaging across repetitions. By normalizing scores, LGN effectively reduces discrepancies between score ranges by narrowing the gaps in score distributions. The general formula for normalization is as follows, with μ and σ being the direction-wise mean and standard deviation:

$$z = \frac{\text{score} - \mu}{\sigma}. \quad (1)$$

6.2 Experiments and Results

We evaluate LGN by applying it before cross-lingual score averaging, following the same experimental setup as in Table 5. Results in Table 7 show that LGN consistently improves the correlation between automatic metrics and human evaluations in multilingual settings. Although the absolute gains are moderate, partly because correlations are already high under the original setup, paired-sample t-tests reported in Appendix K confirm the statistically-significant improvement. Also, this reflects the concern raised in the system-level verification of Section 5.1, where the value shown in Table 5 is high but still suboptimal due to the cross-lingual scoring bias. By reducing disparities in score ranges, LGN improves cross-lingual consistency both the system and triplet levels.[21] This directly addresses the concern raised in the system-level analysis: without normalization, averaging scores across directions is unreliable, as some languages may be systematically over- or under-estimated. Our results suggest that applying LGN before aggregation provides a more reliable basis for multilingual system evaluation. While the generalizability of LGN warrants further investigation, these findings offer initial evidence that normalization-based scoring can mitigate cross-lingual bias in automatic evaluation metrics.

7 Conclusion

In this work, we introduce XQ-MEval, the first multilingual dataset designed to achieve parallel quality across languages for benchmarking automatic evaluation metrics. Based on the benchmark, we identify limitations in the commonly used practice of averaging metric scores across translation directions to represent system-level performance. Specifically, we reveal that cross-lingual scoring bias, caused by metrics exhibiting different scoring ranges across languages, is a key factor contributing to the misalignment between metrics and human evaluation in multilingual settings. Building on this observation, we propose a normalization-based strategy to mitigate cross-lingual scoring bias by narrowing the distances between score ranges. Experimental results show that the LGN strategy significantly improves the consistency with human evaluations and highlight the importance of aligning score ranges across languages to a unified scale before averaging for reliability.

[20] Appendix A describes the computational procedure with pseudo codes.

[21] We also reproduce the analysis in Figure 3 after applying LGN in Appendix L.

Limitations

Human evaluation remains a major bottleneck in machine translation research, as large-scale multilingual annotation, especially for expert-level annotation, is costly and resource-intensive. Although our semi-automatic pipeline alleviates this reliance and makes benchmark construction more efficient, the current version covers only nine translation directions. Nevertheless, the pipeline is highly flexible and can be extended to more languages in future work.

While the MQM framework provides a comprehensive set of error categories, we focus on only four purely semantic error types in our work. However, as discussed in Section 3.1, these error types are better suited for achieving cross-lingual comparability and represent the most prominent categories in existing MQM datasets, accounting for approximately 46.3% of all errors. Although our pipeline can incorporate additional error types, doing so first requires careful linguistic justification to ensure that the added types remain comparable across languages.

Given that this is the first work to discuss the fairness in evaluating multilingual translation systems, our work raises further questions for future research. For instance, are metrics equally sensitive to different error types, or do they respond unevenly? More intriguingly, does this sensitivity vary across languages? We leave these fine-grained investigations for future work.

Ethics Statement

In this work, we construct the XQ-MEval dataset based on Flores, a public dataset, combining manual filtering to enhance its quality. We recruit eligible students from our institution to assist with human annotation tasks, and the compensation provided is in compliance with local standards. All human-involved steps during the construction are carefully designed to ensure that no personal information is involved. The manual annotation process adheres strictly to the ethical guidelines of our institution and the ACL ethics policy. Thus, this recruitment and annotation are approved by the ethics reviewing committee of our affiliation. Generally, this benchmark can be applied in real-world scenarios, supporting the evaluation of automatic evaluation metrics in multilingual settings.

Flores is released under the CC BY-SA 4.0 license[22], which explicitly permits adaptation and sharing. To fully comply with these terms, our license in releasing XQ-MEval would be CC BY-SA 4.0. Moreover, XQ-MEval is created using GPT-4o and is therefore subject to OpenAI's license terms[23]. OpenAI assigns to us all rights, titles, and interests in and to the output.

Use of AI Assistance

During the preparation of this paper, we used ChatGPT to assist with proofreading and polishing. The model was employed solely to improve clarity, grammar, and readability of the manuscript; all ideas, experimental designs, analyses, and conclusions come from the authors. The authors carefully reviewed and verified all AI-assisted edits to ensure correctness and faithfulness to the intended meaning.

References

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454, Bangkok, Thailand. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, and 2 others. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

[22] https://huggingface.co/datasets/openlanguagedata/flores_plus

[23] https://openai.com/policies/terms-of-use

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Zhe Cao, Zhi Qu, Hidetaka Kamigaito, and Taro Watanabe. 2024. Exploring intrinsic language-specific subspaces in fine-tuning multilingual neural machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21142–21157, Miami, Florida, USA. Association for Computational Linguistics.

Liang Chen, Shuming Ma, Dongdong Zhang, Furu Wei, and Baobao Chang. 2023. On the off-target problem of zero-shot multilingual neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9542–9558, Toronto, Canada. Association for Computational Linguistics.

Sonia Cristofaro. 2009. Grammatical categories and relations: universality vs. language-specificity and construction-specificity. *Language and Linguistics Compass*, 3(1):441–479.

Michael Denkowski and Alon Lavie. 2010. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Research Papers*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Salvador García, Julián Luengo, and Francisco Herrera. 2015. *Data Preprocessing in Data Mining*, volume 72.

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

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun

Liu, and Kam-Fai Wong. 2024. MT-eval: A multi-turn capabilities evaluation benchmark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20153–20177, Miami, Florida, USA. Association for Computational Linguistics.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

NLLB Team. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Zhi Qu, Yiran Wang, Jiannan Mao, Chenchen Ding, Hideki Tanaka, Masao Utiyama, and Taro Watanabe. 2025. Registering source tokens to target language spaces in multilingual neural machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21687–21706, Vienna, Austria. Association for Computational Linguistics.

Zhi Qu and Taro Watanabe. 2022. Adapting to non-centered languages for zero-shot multilingual translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5251–5265, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Shaomu Tan and Christof Monz. 2025. ReMedy: Learning machine translation evaluation from human preferences with reward modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4370–4387,

Suzhou, China. Association for Computational Linguistics.

David Vilar, Gregor Leusch, Hermann Ney, and Rafael E. Banchs. 2007. Human evaluation of machine translation through binary system comparisons. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 96–103, Prague, Czech Republic. Association for Computational Linguistics.

Pius Von Däniken, Jan Milan Deriu, and Mark Cieliebak. 2025. A measure of the system dependence of automated metrics. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 87–99, Vienna, Austria. Association for Computational Linguistics.

Weiqi Wang, Limeng Cui, Xin Liu, Sreyashi Nag, Wenju Xu, Chen Luo, Sheikh Muhammad Sarwar, Yang Li, Hansu Gu, Hui Liu, Changlong Yu, Jiaxin Bai, Yifan Gao, Haiyang Zhang, Qi He, Shuiwang Ji, and Yangqiu Song. 2025. EcomScriptBench: A multi-task benchmark for E-commerce script planning via step-wise intention-driven product association. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–22, Vienna, Austria. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. 2021. Share or not? learning to schedule language-specific capacity for multilingual translation. In *International Conference on Learning Representations*.

A Computational Procedure

Algorithm 1 formalizes the average strategy described in Section 1, which evaluates multilingual MT systems by first computing metric scores for each triplet (Step 5) and then averaging scores across all translation directions to obtain a system-level score (Step 14). Two highlighted components further clarify key aspects of our evaluation setup. Step 15 computes the corresponding human score to serve as the predefined performance used to benchmark metrics against human judgments, as discussed in Section 3.3. In addition, Step 7 present the normalization based LGN strategy proposed in Section 6.1, where triplet level metric scores are normalized after computation.

Algorithm 1 Evaluation with Average Strategy

1: **Input:** number of language pairs N; number of triplets per language pair I; metric scoring function Metric($\tilde{r}$); human scoring function Human($\tilde{r}$); normalization flag USE_LGN; normalization function LGN(s_m)
2: **Output:** overall metric score S_M; overall human score S_H
3: **for** $i \leftarrow 1$ to N **do** ▷ language pairs
4: **for** $j \leftarrow 1$ to I **do** ▷ triplets
5: $s_m^{(j)} \leftarrow \text{Metric}(\tilde{r}_{i,j})$
6: **if** USE_LGN **then**
7: $s_m^{(j)} \leftarrow \text{LGN}(s_m^{(j)})$
8: **end if**
9: $s_h^{(j)} \leftarrow \text{Human}(\tilde{r}_{i,j})$
10: **end for**
11: $\bar{s}_m^{(i)} \leftarrow \frac{1}{I}\sum_{j=1}^{I} s_m^{(j)}$
12: $\bar{s}_h^{(i)} \leftarrow \frac{1}{I}\sum_{j=1}^{I} s_h^{(j)}$
13: **end for**
14: $S_M \leftarrow \frac{1}{N}\sum_{i=1}^{N} \bar{s}_m^{(i)}$
15: $S_H \leftarrow \frac{1}{N}\sum_{i=1}^{N} \bar{s}_h^{(i)}$
16: **return** S_M, S_H

B Language Selection in Benchmark Construction

In constructing the benchmark, we select nine target languages paired with English, resulting in nine translation directions: `en-zh`, `en-lo`, `en-ja`, `en-vi`, `en-id`, `en-es`, `en-fr`, `en-si`, and `en-de`. This selection aims to ensure a comprehensive evaluation across high-resource and low-resource languages. As discussed in Section 2, most widely-used metrics are driven by MQM-style training, i.e., fine-tuned on MQM-annotated data. However, MQM annotations are only available for high-resource languages, resulting in an imbalanced data distribution. Intuitively, this imbalance may lead MQM-driven metrics to exhibit stronger biases when evaluating translations in low-resource languages. In addition, practical constraints such as the availability of native-speaking volunteers for filtering pseudo translations also influence our language choices. Taking these factors into account, we determine that the selected translation directions strike a reasonable balance between linguistic diversity and feasibility, making the benchmark both representative and manageable. In addition, among the selected languages, those supported by MQM training data include: `zh`, `de`, `es`,

ja, and fr; Languages without MQM support include: lo, vi, id, and si.

C Preliminary Analysis on MQM

As mentioned in Section 1, investigating cross-lingual scoring bias requires instances with strictly parallel semantics and quality. However, MQM datasets cover only a limited number of language pairs, among which only en-de and en-ru satisfy this requirement. For these two directions, we partition instances into five MQM score ranges: 0, $(0, 5]$, $(5, 10]$, $(10, 15]$, and $(15, 25]$, merging the highest range due to data sparsity. We evaluate these instances using BLEURT, XCOMET, and COMETKIWI-23 (spanning all metric types).

The results in Figure 5 show that translations of comparable quality in different language pairs are assigned different scores by the metrics, particularly XCOMET, even when only two language pairs are involved. Moreover, the results demonstrate that cross-lingual scoring bias exists in MQM data and follows a trend similar to that observed in Figure 3, thereby validating our synthetic instances.

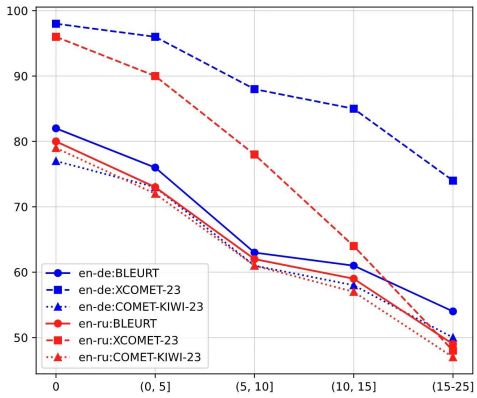

Figure 5: Visualization of three metrics scores across two directions at varying translation quality levels on MQM dataset.

D Annotation Guidelines

To ensure that native speakers acquire a clear understanding of the purpose of our experiment and the definition of MQM, thereby enabling them to more accurately identify and filter error candidates that meet the required criteria, we comply an instruction document that provides the necessary background information and operational guidelines. It is included in the following:

Background

In the evaluation of translation quality, a human-centric framework known as **Multidimensional Quality Metrics (MQM)** (https://themqm.org/) is widely used. Specifically, MQM classifies translation quality based on a standardized error taxonomy, resulting in a scoring system that is both low in subjectivity and high in comparability. This framework significantly facilitates both production and research efforts.

However, MQM annotation is inherently inefficient and costly, as it heavily depends on the manual work of expert annotators. While, in theory, advanced artificial intelligence could act as expert-level annotators, such a substitution is not entirely trustworthy because we cannot verify whether the AI has truly reached expert proficiency.

Fortunately, and interestingly, our task is **NOT** to evaluate a machine translation system in the MQM style. Instead, we aim to obtain MQM-style scores. Specifically, this means we can use advanced AI systems to disrupt a set of perfect translations by introducing errors defined under MQM. Then, we simply ask native speakers to verify whether the disruption was successful. This approach allows us to obtain reliable MQM scores on a given dataset.

Task

Each volunteer will be provided with four files, named en-{lang}-{error}.tsv, where {lang} points to each volunteer's native language, and {error} refers to four common and easily quantified types of errors in machine translation: *Addition*, *Omission*, *Mistranslation*, and *Untranslated*.

In each file, there are three parts that should be noticed:

- **src**: The source sentence in English.
- **ref**: The correct (perfect) translation of the source sentence in the volunteer's native language.
- **mt**: The sentence that has been disrupted by using GPT-4o. Specifically, GPT-4o introduced an error into each ref.

Please note, the error in mt is marked by <v> </v>. Now, you should check the quality of mt, and judge whether the error marked by <v> </v> indeed disrupts ref without any change in the rest

part of `ref`. If the answer is **YES**, you don't need to take any action; otherwise, you should write `T` in the reject column to indicate that the disruption is not acceptable.

Criteria

The following are the evaluation criteria for each type of error:

Addition

The error in `mt` marked by `<v></v>` introduced additional semantics into `ref`.

- If the error indeed presents additional semantics in the `ref` without any change in the rest part of `ref`, then this `mt` is acceptable, i.e., you don't need to take any action.
- Otherwise, please write `T` in the reject column.
- Note that the key is whether the semantics are changed, i.e., a change in the adverb of degree is considered as reject.

Omission

The `mt` has a missing part compared to the `ref`, and the missing part is marked by `<v></v>`.

- If the missing part in the `mt` causes a change in meaning, then this `mt` is acceptable, i.e., you don't need to take any action.
- Note, omission could make the sentence unreadable. However, the unique criterion is that the part outside of labeling (`<v></v>`) is not changed.
- In languages using spaces as intervals, some words could be labeled. However, the following case caused by changes of punctuation is also acceptable:

  > `ref`: ..., en particulier les affaires de voitures volées, avec l'intention...
  > `mt`: ..., en <v>particulier,</v> avec l'intention...

  Here, `les affaires de voitures volées` is omissive, and the label is caused by the change of `particulier` → `particulier,`.
- Otherwise, e.g., the missing part changes the part out of `<v></v>` or the marked part is not missed, please write `T` in the reject column.

Mistranslation

The error in `mt` marked by `<v></v>` is a mistranslation from `src`.

- Given that `ref` is a ground-truth translation from `src`, you can simply compare `ref` and `mt`. If the error of `mt` conveys different words or semantics compared to `ref`, this `mt` is acceptable, i.e., you don't need to take any action.
- Otherwise, please write `T` in the reject column.

Untranslated

The error in `mt` marked by `<v></v>` has not been translated and remains in the original English.

- Simply copying from `src` or changing words but remaining in English is recognized as acceptable, i.e., you don't need to take any action.
- If the untranslated words are person's names or place names, please write `T` in the reject column.

Overall

Changes in the content of the `mt` may result in grammatical errors in the overall sentence, and this is acceptable as long as the part marked with `<v></v>` in the `mt` indeed causes a change in meaning without changes in the part outside of `<v></v>`. This indicates that the `mt` is acceptable.

E Prompt Design

To instruct GPT-4o to introduce addition, omission, mistranslation and untranslated errors to references to obtain temporary error candidates containing one error segment, we design the specific prompt for different error types. Figure 6 shows the details of the prompt.

F Triplets Count Distribution

Table 8 shows the triplets count distribution across the five quality levels for each language pair. As shown in the table, the triplets with quality level 2 and 3 are more frequent, while triplets at level 5 are fewer. This is because quality level reflects the number of errors in pseudo translations; as the error count increases, overlapping error spans reduce the number of generated triplets.

(System) Your task is to introduce errors to disrupt the quality.

"Addition": \n
Given a sentence, your task is to add an addition error to disrupt the quality.
Please do the following instructions step by step:
1. You should select a sub-part of the sentence in the part enclosed by <{position}> and </{position}>, then output the sub-part you selected.
2. You should disrupt the sub-part by adding some words which includes information not present in the selected sub-part. Please keep the rest the same. Then output the disrupted sub-part.
3. Replace the selected sub-part by the disrupted sub-part to get the updated sentence.
4. Finally, output the updated sentence.

"Omission": \n
Given a sentence, your task is to add an omission error to disrupt the quality.
Please do the following instructions step by step:
1. You should select a sub-part of the sentence in the part enclosed by <{position}> and </{position}>, then output the sub-part you selected.
2. You should select a segment containing some important information in the sub-part. Note that a segment means some words or a phrase rather than a clause. Then output the segment you selected.
3. You should delete the segment from the sub-part to get the disrupted sub-part, make sure that you just delete one segment.
4. Replace the sub-part in the sentence by the disrupted sub-part to get the updated sentence.
5. Finally, output the updated sentence.

"Mistranslation": \n
Given a sentence, your task is to add an mistranslation error to disrupt the quality.
Please do the following instructions step by step:
1. You should select a sub-part of the sentence in the part enclosed by <{position}> and </{position}>, then output the sub-part you selected.
2. You should select a segment containing some important information in the sub-part. Ensure the segment is a natural and coherent phrase rather than fragments of different sentences or clauses. And the segment is typically a short phrase that conveys a key idea without unnecessary details. Then output the segment you selected.
3. You should replace the segment you selected in the sub-part, with alternatives that change the meaning of that part to get the disrupted segment. Do NOT perform simple substitutions, such as replacing "1" with "2" or "good" with "bad". Use descriptive phrases or reframe the meaning to introduce different information. Then output the disrupted segment.
4. Replace the selected segment in the selected sub-part by the disrupted segment to get the disrupted sub-part, then output the disrupted sub-part.
5. Replace the selected sub-part in the sentence by the disrupted sub-part to get the updated sentence.
6. Finally, output the updated sentence.

"Untranslated": \n
Given a source sentence and target sentence, your task is to add an untranslation error to disrupt the translation quality.
Please do the following instructions step by step:
1. You should select a sub-part of the target sentence in the part enclosed by <{position}> and </{position}>.
Note that, a sub-part means a word or a phrase instead of a clause. Then output the sub-part you selected.
2. Given that our objective is to create an untranslation error, the selected sub-part should be in {language} instead of English and does not present in the source sentence. Please validate it. If it cannot meet our requirement, please select another sub-part in {language}.
3. You should find the corresponding part from the source sentence. Then output the corresponding source part.
4. Replace the selected sub-part by the corresponding source part to get the updated target sentence, finally output the updated sentence.

Figure 6: The prompt for different error types to guide GPT-4o to introduce errors to references.

G Discussion on Repeated Sampling

To examine the effect of repeated sampling on evaluation stability, we test three metrics, i.e., BLEURT, xCOMET, and KIWI23, on monolingual systems for `en-zh`, `en-ja`, and `en-de` at five quality levels. For each system, 102 triplets are sampled, and the procedure is repeated 5, 10, and 25 times. Table 10 reports the means and variances across these settings. As the sampling iterations increase, the mean scores shown in the table exhibit stability. Although the variance fluctuates to some extent, it is caused by the value is scaled to the square of the scoring scale because the scores are amplified by a factor of 100. Consequently,

Quality Level	en-zh	en-lo	en-ja	en-vi	en-id	en-fr	en-es	en-si	en-de
1	776	753	775	771	782	775	771	765	774
2	2,109	2,053	2,078	2,056	2,095	1,992	2,016	2,064	2,049
3	2,548	2,627	2,441	2,420	2,421	2,068	2,233	2,489	2,337
4	1,466	1,704	1,324	1,387	1,311	9,57	1,069	1,432	1,234
5	406	558	340	428	312	198	203	361	313

Table 8: The triplets count distribution across the five quality levels for each language pair.

Num.	System-level	Triplet-level
3	0.03	0.03
6	0.009	0.003
9	0.001	0.005

Table 9: Paired samples t-test results for system-level and triplet-level improvements obtained with the LGN strategy.

the variance remains within a small range, and we consider these fluctuations to be acceptable. Ultimately, we adopt the approach of repeating the process 10 times in our main experiments.

H Detailed Scores

Table 11 presents the detailed scores of Figure 3.

I Score Reduction across Directions and Metrics

Figure 3 reveals that as translation quality declines, the rate of score reduction differs across translation directions, highlighting the varying sensitivities of metrics to quality changes across languages. This variation exacerbates score inconsistencies across directions at the same quality level, particularly for lower-quality translations. The widening gaps in the decline patterns further illustrate this trend. Similarly, score reduction patterns differ across metrics. For spBLEU, scores are approaching at high quality but diverge as quality decreases due to different decline rates across directions. chrF shows more consistent decline trends, though its score ranges vary substantially across directions, with `zh` and `ja` exhibiting systematically lower ranges. BLEURT exhibits behavior similar to spBLEU, but with larger cross-lingual discrepancies in score reduction as translation quality deteriorates. For COMET and xCOMET, score reduction trends exhibit similar patterns across directions. However, COMET assigns direction-specific score ranges with limited overlap, whereas xCOMET produces more aligned score ranges for most directions, except `lo`, `si`, and `de`. In contrast, KIWI22 and KIWI23 more closely align with the desired properties of an ideal metric, as they exhibit more closely aligned score ranges and score reduction trends, whereas KIWI23 still shows noticeable score range discrepancies for certain directions. By comparison, the MetricX variants display substantial cross-lingual inconsistency in both score ranges and reduction patterns, with regression-based MetricX exhibiting pronounced inconsistencies.

J Experiment on LLM-based Metrics

We investigate two LLM-based evaluation approaches: ReMedy (Tan and Monz, 2025), a trainable evaluation metric fine-tuned from LLMs; and GEMBA-MQM (Kocmi and Federmann, 2023), which prompts LLMs to simulate human annotators by following MQM guidelines. Using these evaluators, we assess translation triplets at varying quality levels across three language pairs, `en-zh`, `en-es`, and `en-ja`, reflecting the language coverage of ReMedy in our study. The results in Figure 7 reveal substantial variation across language pairs, indicating that LLM-based evaluators remain susceptible to cross-lingual bias.

K Significance Test

We conduct paired samples t-tests on the improvements obtained with the LGN strategy in Table 7. As shown in the Table 9, all p-values are below 0.05, indicating that although the improvements are small in magnitude, they are statistically significant.

L Results under the LGN Strategy

Figure 8 shows the normalized scores of nine metrics across translation directions at varying quality levels. As illustrated in the figure, the LGN strategy effectively narrows score range disparities across language pairs, as evidenced by the reduced distances between curves. After applying LGN, translations of comparable quality from different language pairs receive consistent metric scores, and the score degradation trends as translation quality decreases become more consistent across directions.

	BLEURT					xCOMET					KIWI23				
mean	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
en-zh	81.98	70.37	62.36	56.45	51.85	82.90	68.44	55.50	45.20	36.66	66.38	56.09	46.43	40.23	33.39
en-ja	81.65	69.15	60.18	53.47	48.18	80.85	63.93	49.20	38.78	32.51	67.15	57.28	49.18	43.50	39.67
en-de	80.05	67.82	59.01	52.02	47.85	91.89	83.21	72.83	64.83	56.65	61.92	49.16	38.19	29.52	25.43
var	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
en-zh	0.36	0.36	1.49	0.40	0.62	0.55	0.85	1.27	1.21	0.74	0.33	0.69	1.04	0.07	1.68
en-ja	0.41	0.58	0.10	0.80	0.83	0.35	2.22	2.99	3.39	1.60	0.27	1.44	0.89	0.40	1.43
en-de	0.39	1.94	1.22	1.83	0.37	0.19	0.26	2.71	2.57	4.01	0.94	1.38	3.70	1.43	1.68

(a) Mean and variance for 5 iterations of sampling.

	BLEURT					xCOMET					KIWI23				
mean	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
en-zh	81.90	70.60	62.42	56.36	52.09	82.93	68.49	56.06	45.32	37.41	66.28	56.12	46.95	39.87	33.83
en-ja	81.97	69.27	60.36	53.74	48.21	81.12	64.38	48.79	38.62	32.48	67.66	57.38	49.36	43.73	39.54
en-de	79.63	67.99	59.62	52.23	47.35	91.45	83.20	73.64	64.12	55.40	60.92	48.82	38.95	29.87	24.84
var	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
en-zh	0.22	0.42	1.00	0.36	0.62	0.67	0.60	2.36	0.85	1.02	0.29	0.41	2.05	0.17	1.66
en-ja	0.45	0.43	0.36	0.89	0.58	0.79	1.56	2.26	2.87	1.28	0.72	0.76	0.69	0.59	0.94
en-de	0.71	1.33	1.52	1.18	0.64	0.80	0.57	4.47	2.24	6.12	1.84	1.54	4.35	0.99	1.95

(b) Mean and variance for 10 iterations of sampling.

	BLEURT					xCOMET					KIWI23				
mean	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
en-zh	81.80	70.52	62.39	56.21	51.91	82.56	68.64	55.88	45.51	37.23	66.10	55.66	46.72	39.93	33.59
en-ja	81.98	69.71	60.58	53.65	47.92	81.34	64.51	49.80	38.56	32.04	67.72	57.55	49.73	43.68	39.23
en-de	79.63	68.04	59.47	52.91	47.29	91.41	83.23	73.69	64.79	55.16	60.89	48.54	38.51	30.84	24.78
var	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
en-zh	0.34	0.53	1.07	0.64	0.78	0.77	1.52	2.67	1.37	0.88	0.43	1.49	1.82	1.01	1.10
en-ja	0.42	0.49	0.81	0.73	0.71	1.87	1.86	3.87	2.37	1.77	0.90	0.56	1.89	0.79	1.25
en-de	0.50	0.99	1.25	1.49	0.78	0.59	1.01	2.48	3.04	3.75	1.49	1.92	3.15	2.95	1.31

(c) Mean and variance for 25 iterations of sampling.

Table 10: Mean and variance for 5, 10, 25 iterations of sampling. Note that the scores are amplified by a factor of 100, and the scale of the variance corresponds to the square of the scoring scale.

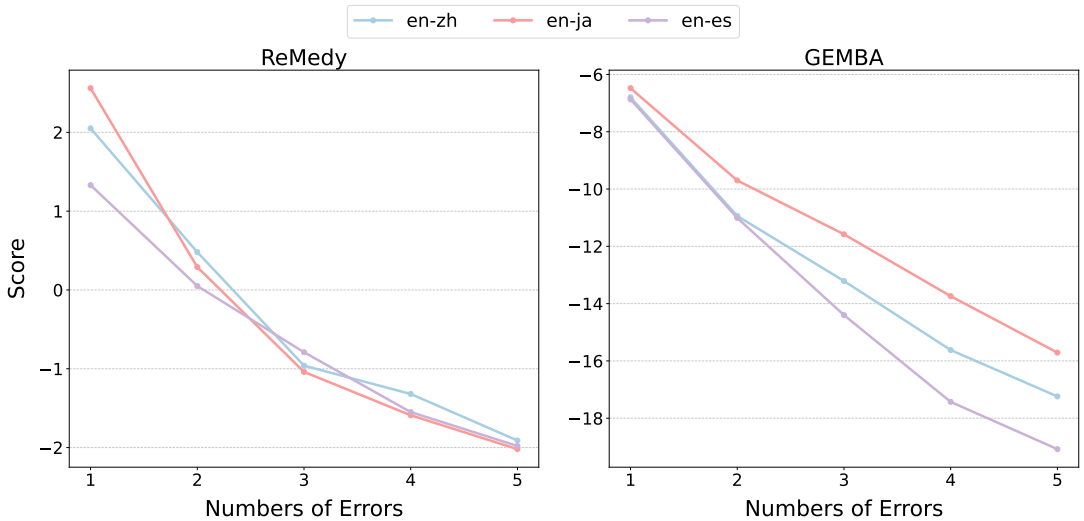

Figure 7: Visualization of LLM-based evaluation scores across three directions at varying translation quality levels.

	spBLEU					chrF					BLEURT				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
en-zh	83.85	70.50	60.36	51.84	45.16	74.46	62.94	54.34	46.97	41.70	81.90	70.81	62.60	56.89	52.61
en-lo	84.99	73.74	64.88	55.64	48.21	87.33	77.31	69.93	63.03	57.35	82.87	73.03	65.26	57.72	52.88
en-ja	81.13	66.94	55.20	45.00	35.78	75.03	65.13	57.05	50.47	44.38	81.97	69.74	61.05	54.29	47.59
en-vi	84.21	72.71	62.93	56.48	51.55	90.44	82.54	75.98	71.39	67.88	81.85	71.86	63.54	57.29	52.25
en-id	83.10	70.02	60.12	51.25	44.98	90.83	82.89	76.89	71.49	67.46	83.15	74.56	68.42	64.55	62.11
en-fr	84.03	71.60	62.46	55.21	48.60	90.41	82.03	75.74	70.37	65.28	80.96	67.37	56.23	47.83	40.07
en-es	84.30	71.72	62.35	53.94	46.73	90.80	82.63	76.49	71.52	67.16	83.37	71.68	62.18	54.59	48.96
en-si	86.18	75.42	67.07	59.80	53.27	91.40	83.41	77.29	71.65	66.73	84.38	78.44	72.83	68.57	64.73
en-de	84.09	72.25	62.61	54.31	48.01	90.93	83.13	76.58	71.38	66.47	79.63	68.03	59.56	53.64	46.88

(a) Sequenced-based metrics.

	COMET					xCOMET					MX-reg				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
en-zh	90.42	85.97	82.17	78.46	75.10	82.93	68.65	55.82	45.86	37.67	2.40	3.58	4.75	5.63	6.86
en-lo	88.15	83.31	79.07	74.60	71.16	63.72	48.64	36.11	27.04	21.01	3.46	5.49	7.24	8.81	10.22
en-ja	92.63	88.90	85.18	81.85	77.12	81.12	64.52	49.83	39.70	31.86	1.85	2.91	3.87	4.91	6.03
en-vi	89.81	85.61	81.59	78.68	75.26	81.83	68.44	55.56	46.71	36.54	3.22	5.05	6.55	8.05	9.71
en-id	91.72	87.93	83.95	80.52	77.35	84.43	70.58	55.69	45.30	36.19	2.52	3.89	5.53	6.65	7.58
en-fr	86.49	80.36	74.45	69.70	64.79	82.18	66.13	51.27	40.37	30.00	2.64	4.04	5.51	6.81	8.40
en-es	87.15	81.37	75.73	70.81	66.54	86.06	72.62	59.45	47.80	37.82	2.44	3.90	5.40	6.37	7.28
en-si	92.51	89.30	85.92	82.84	79.64	69.67	52.42	37.07	26.85	20.31	2.30	4.08	6.08	7.67	9.83
en-de	87.39	81.18	75.75	70.89	64.36	91.45	83.13	73.95	66.14	55.02	1.53	2.23	2.87	3.40	4.17

(b) Regression-based metrics.

	KIWI22					KIWI23					MX-qe				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
en-zh	76.40	69.10	63.34	58.68	54.74	66.28	55.69	46.63	40.37	34.33	1.70	2.32	3.00	3.54	4.26
en-lo	74.50	68.24	63.82	60.12	57.88	61.41	52.93	46.59	40.60	36.18	2.71	3.87	4.83	5.81	6.62
en-ja	79.37	72.80	67.69	63.90	60.15	67.66	57.27	49.42	44.60	38.96	1.38	2.04	2.65	3.25	4.19
en-vi	76.38	70.37	65.30	61.81	58.60	64.94	55.71	48.44	43.27	38.43	2.04	2.89	3.62	4.43	5.11
en-id	76.09	69.15	63.20	58.60	55.54	66.37	56.34	47.66	41.96	37.25	2.23	3.12	4.16	4.94	5.54
en-fr	78.60	70.98	64.87	60.42	55.42	60.82	47.49	38.08	30.50	24.20	1.97	2.83	3.85	4.70	5.94
en-es	76.16	68.48	61.98	57.04	53.43	64.96	53.30	43.91	36.34	30.66	1.97	2.84	3.87	4.63	5.26
en-si	80.42	72.71	65.64	60.72	56.20	74.13	64.18	55.82	49.46	44.87	1.25	1.89	2.74	3.72	4.93
en-de	75.62	68.56	62.28	58.17	53.91	60.92	48.52	38.47	31.73	24.32	1.44	2.11	2.71	3.35	4.21

(c) Regression-free metrics.

Table 11: The detailed scores of nine metrics when evaluating different languages at various quality levels.

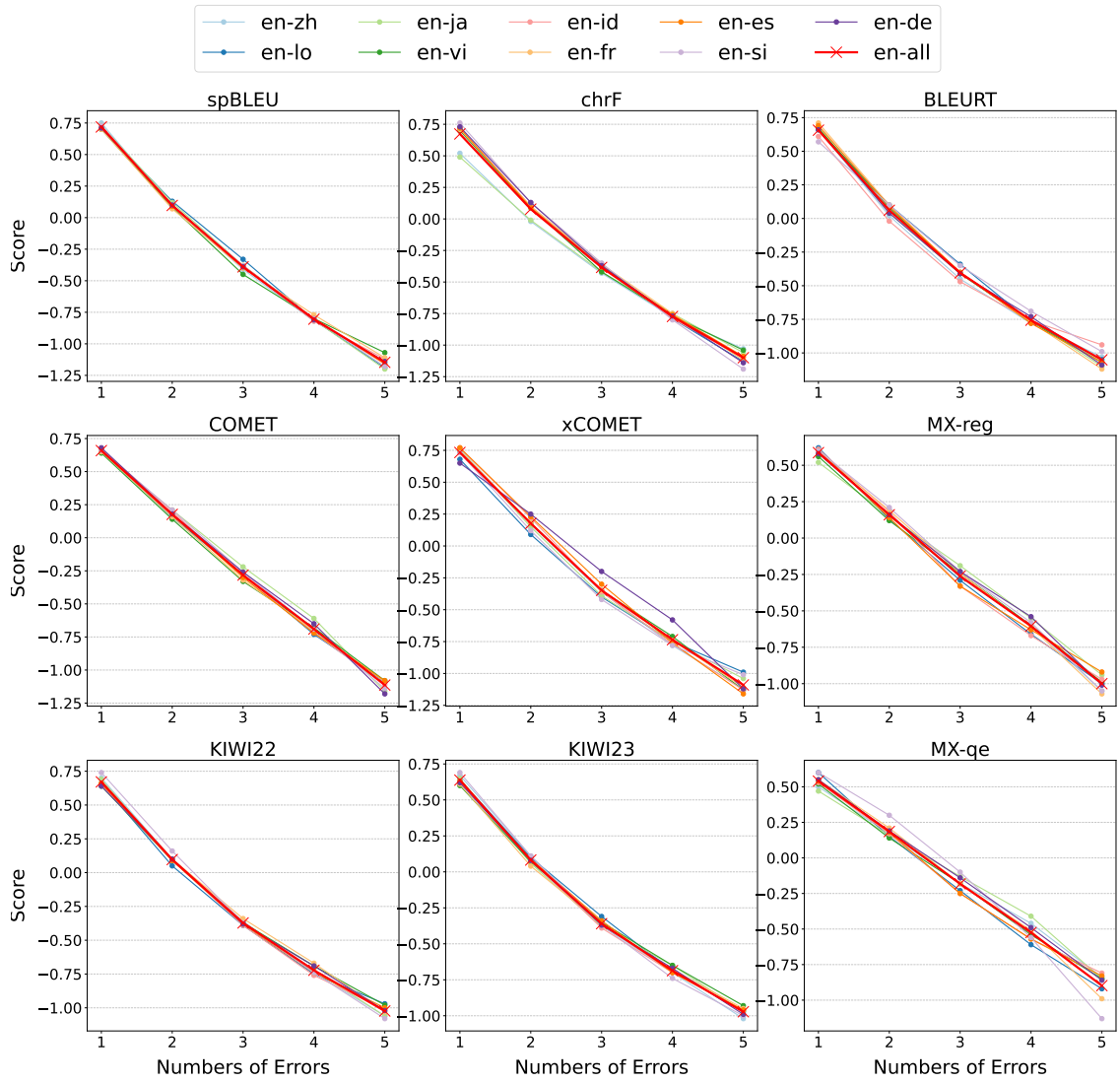

Figure 8: Visualization of nine metrics scores under the LGN strategy across nine directions at varying translation quality levels. `en-all` denoting the average metric scores among all directions.