Beyond Correlation: Interpretable Evaluation of Machine Translation Metrics

Stefano Perrella* **Lorenzo Proietti*** **Pere-Lluís Huguet Cabot**
Edoardo Barba **Roberto Navigli**
Sapienza NLP Group, Sapienza University of Rome
{perrella, lproietti, huguetcabot, barba, navigli}@diag.uniroma1.it

Abstract

Machine Translation (MT) evaluation metrics assess translation quality automatically. Recently, researchers have employed MT metrics for various new use cases, such as data filtering and translation re-ranking. However, most MT metrics return assessments as scalar scores that are difficult to interpret, posing a challenge to making informed design choices. Moreover, MT metrics' capabilities have historically been evaluated using correlation with human judgment, which, despite its efficacy, falls short of providing intuitive insights into metric performance, especially in terms of new metric use cases. To address these issues, we introduce an interpretable evaluation framework for MT metrics. Within this framework, we evaluate metrics in two scenarios that serve as proxies for the data filtering and translation re-ranking use cases. Furthermore, by measuring the performance of MT metrics using Precision, Recall, and F-score, we offer clearer insights into their capabilities than correlation with human judgments. Finally, we raise concerns regarding the reliability of manually curated data following the Direct Assessments+Scalar Quality Metrics (DA+SQM) guidelines, reporting a notably low agreement with Multidimensional Quality Metrics (MQM) annotations.

1 Introduction

Over the past few years, Machine Translation (MT) evaluation metrics have transitioned from heuristic-based to neural-based, enabling a more nuanced evaluation of translation quality and a greater agreement with human judgments (Freitag et al., 2022b). Additionally, recent Metrics Shared Tasks at the Conference on Machine Translation (Mathur et al., 2020b; Freitag et al., 2021b, WMT) have seen the rise of reference-free metrics, which assess translation quality without the need for human-curated references by comparing translations only to their sources in the original language. Lately, reference-free metrics have demonstrated performance on par with, and sometimes superior to, their reference-based counterparts (Freitag et al., 2023; Kocmi et al., 2024a). Thanks to these advancements and the ability to use metrics without references, several new MT metrics use cases have emerged. Freitag et al. (2022a), Fernandes et al. (2022), Farinhas et al. (2023), Ramos et al. (2024), and Finkelstein and Freitag (2024) used MT metrics as utility functions for Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004; Eikema and Aziz, 2020) and for Quality Estimation (QE) re-ranking.[1] Ramos et al. (2024), Gulcehre et al. (2023), He et al. (2024), and Xu et al. (2024) used MT metrics as a proxy for human preferences to fine-tune MT systems using Reinforcement Learning (RL)- and Direct Preference Optimization (DPO)-like training objectives. Peter et al. (2023), Alves et al.

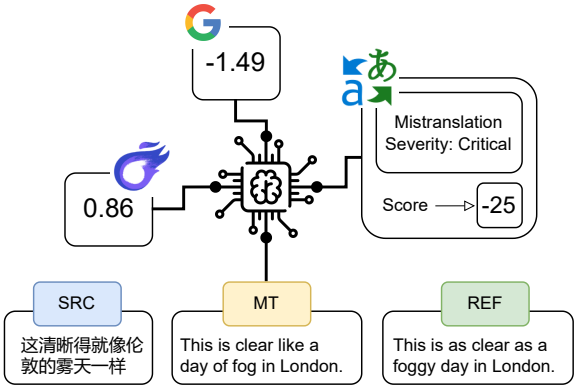

Figure 1: Quality assessments returned by COMET (Rei et al., 2020), MetricX-23-QE-XL (Juraska et al., 2023), and GEMBA-MQM (Kocmi and Federmann, 2023) for the provided machine-translated text.

*Equal contribution.

[1]MBR decoding and QE re-ranking are methods used to identify the best translation from multiple outputs generated by an MT system for the same source text. MBR decoding typically relies on reference-based metrics, while QE re-ranking depends on reference-free metrics.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 20689–20714
November 12-16, 2024

(2024), and Gulcehre et al. (2023) used reference-free metrics to filter parallel corpora – discarding all translations assigned with a metric score that is below a certain threshold – with the goal of training MT systems using higher quality data. These works leverage MT metrics for applications beyond their traditional use of measuring incremental improvements in the development of MT systems. However, the lack of a dedicated evaluation, paired with the inherent opacity of MT metrics, makes it challenging to determine whether one metric suits a given task better and what the impact of various design choices is. For example, Alves et al. (2024), Peter et al. (2023), and Gulcehre et al. (2023) filter MT datasets using different MT metrics and thresholds, leaving it unclear whether an optimal choice exists. Furthermore, considering the ever-increasing number of metrics available, researchers are often limited to grid-searching for the best configuration for each new application, as do Fernandes et al. (2022) and Ramos et al. (2024), who explore by grid-search whether certain metrics are better suited than others for MBR decoding, QE re-ranking, and as reward models for RL-based training. However, the lack of dedicated evaluation setups often requires revisiting these studies to assess whether their findings hold with newer metrics, resulting in a non-negligible increase in experimentation time.

In this work, we address these issues by introducing a novel and more interpretable evaluation framework for MT metrics, comprising evaluation setups designed as proxies for new metric use cases. In the following sections, we first illustrate the problem of interpretability, then introduce our framework, and finally present our results.

2 The Interpretability of MT Metrics' Assessments

In the field of AI, Interpretability is defined as "the ability to explain or to provide the meaning in understandable terms to a human" (Barredo Arrieta et al., 2020), and typically refers to the problem of understanding the decision-making process of an AI model. However, our goal here is less ambitious. Instead of focusing on the interpretability of MT metrics themselves, we are concerned with the interpretability of their assessments. Specifically, most state-of-the-art MT metrics are trained to minimize the Mean Squared Error (MSE) with human judgments and return assessments as scalar quality scores, which are difficult to interpret. Therefore, we are interested in understanding the meaning of these scores, rather than the internal workings of MT metrics.

In light of this, we attribute MT metrics assessments' lack of interpretability to three main factors:[2]

1. **Range consistency**: it is unclear whether a difference in metric score has the same meaning if it occurs in different regions of the score range.

2. **Error attribution**: scalar quality assessments do not identify specific translation errors.

3. **Performance**: metrics capabilities are typically measured through correlation with human judgment, which fails to provide users with a clear understanding of their performance and reliability.

In simpler terms, let us consider the example of Figure 1. Due to the lack of *Error attribution* we do not know which translation errors, if any, led COMET (Rei et al., 2020) to return 0.86. Also, the metric comes with no *Range consistency* guarantee, e.g. whether 0.86 is twice as good as 0.43. Furthermore, different metrics have different score ranges, making it difficult to compare the assessments from COMET with those of the other MT metrics in the figure. Finally, lacking a clear understanding of COMET's *Performance* beyond human correlation, we cannot be sure whether we can draw conclusions from its assessments safely.

For this reason, some efforts have been made to design interpretable metrics. For example, among the primary submissions to the WMT23 Metrics Shared Task (Freitag et al., 2023), MaTESe (Perrella et al., 2022) annotates the spans of a translation that contain errors, specifying their severity, xCOMET models (Guerreiro et al., 2024) return annotated error spans together with a final regression value, and GEMBA-MQM (Kocmi and Federmann, 2023) leverages GPT-4 (OpenAI et al., 2024) to produce detailed quality assessments. However, these metrics compromise on other aspects to accommodate the increased interpretability. MaTESe displays a lower correlation with human judgment

[2]We wish to clarify that we identified these three factors as notably impactful, but they are non-exhaustive and may overlap.

than several regression-based metrics.[3] Trading off performance and interpretability, XCOMET models' final assessment is based mainly on the regression component, with annotated spans contributing only $2/9$ of the overall score. Finally, GEMBA-MQM is prohibitively expensive to operate and its assessments are not fully-reproducible, due to its dependence on the closed-source GPT-4.

In this work, we seek to mitigate the interpretability issue by targeting the problem of *Performance*. Specifically, we take inspiration from two popular new MT metrics applications, i.e., data filtering and translation re-ranking, in order to study and measure metrics performance in terms of Precision, Recall, and F-score. Taking advantage of these measures that are more transparent than correlation, we aim to shed light on the meaning and reliability of metrics assessments, especially concerning such new MT metrics use cases. We release our evaluation framework as software at https://github.com/SapienzaNLP/interpretable-mt-metrics-eval.

3 An Interpretable Evaluation Framework for MT Metrics

Two popular new MT metrics applications are data filtering and translation re-ranking. In data filtering, MT metrics separate good-quality from poor-quality translations. After choosing a threshold value, all translations below the threshold are labeled as poor quality and discarded. In this respect, we are interested in jointly assessing metric performance and studying the meaning of metric scores, finding the thresholds that best separate good-quality (GOOD) from poor-quality (BAD) translations. Instead, in translation re-ranking, MT metrics determine the best in a pool of translations of the same source text. For example, in QE reranking and MBR decoding, metrics are tasked to identify the best translation among those sampled from an MT system.

With the aim of facilitating practitioners in making design choices for these metrics applications, and with a focus on the interpretability issue, we evaluate MT metrics performance in two settings: i) when metrics are used as binary classifiers, tasked to separate between GOOD and BAD translations (acting as a proxy for the data filtering application), and ii) when metrics are used to identify the best translation in a group of translations of the same source (acting as a proxy for translation reranking).

3.1 Metrics as Binary Classifiers for Data Filtering

Let us consider the MT metric $\mathcal{M}$, which outputs scores in the range $[m_1, m_2]$. Let us define $\mathcal{M}(t) \in [m_1, m_2]$ as the score assigned by metric $\mathcal{M}$ to translation t. By selecting an arbitrary threshold value $\tau \in [m_1, m_2]$, we repurpose $\mathcal{M}$ as a binary classifier: a translation t is deemed as GOOD by metric $\mathcal{M}$, with threshold τ, if $\mathcal{M}(t) \geq \tau$, otherwise it is deemed as BAD .

Precision, Recall, and *F*-score Assuming that we have an oracle $\mathcal{H}$ telling us whether a translation is GOOD or BAD , we can measure the performance of metric $\mathcal{M}$, with threshold τ, in terms of standard measures such as Precision, Recall, and F-score, which we refer to as $P^{\mathcal{M}_\tau}$, $R^{\mathcal{M}_\tau}$, and $F^{\mathcal{M}_\tau}$. Given metric $\mathcal{M}$, oracle $\mathcal{H}$, translation t, and threshold τ, $P^{\mathcal{M}_\tau}$ estimates the probability that translation t is GOOD , given that metric $\mathcal{M}$ deems it as such:

$$P^{\mathcal{M}_\tau} = \hat{\Pr}(\mathcal{H}(t) = \text{GOOD} \mid \mathcal{M}(t) \geq \tau). \quad (1)$$

Similarly, $R^{\mathcal{M}_\tau}$ estimates the probability that metric $\mathcal{M}$ deems translation t as GOOD , given that the oracle deems it as such:

$$R^{\mathcal{M}_\tau} = \hat{\Pr}(\mathcal{M}(t) \geq \tau \mid \mathcal{H}(t) = \text{GOOD}). \quad (2)$$

Finally, we aggregate Precision and Recall using F_β-score, with $\beta = \frac{1}{\sqrt{2}}$, which weights Precision higher than Recall compared to the more common F_1-score. Arguably, false positives – i.e., translations of low quality that are mistakenly considered GOOD – could be detrimental to the applications that see metrics employed as binary classifiers. For example, in data filtering, false positives correspond to low-quality translations that survive the filtering, compromising the quality of filtered data. In contrast, false negatives – i.e., translations of high quality that are mistakenly assigned with the BAD label – would more frequently lead to minor inconveniences, as they correspond to good-quality translations that are mistakenly discarded. Moreover, we note that MT metrics struggle to achieve high Precision, meaning that metrics differences can be best highlighted if Precision is weighted higher than Recall.

[3] While this might be due to several contributing factors, the limited availability of training data containing detailed span-level annotations is most likely one of them.

Therefore, the F-score of metric $\mathcal{M}$, with threshold τ, is defined as follows:

$$F^{\mathcal{M}_\tau} = \frac{3}{2} \frac{P^{\mathcal{M}_\tau} R^{\mathcal{M}_\tau}}{\frac{1}{2} P^{\mathcal{M}_\tau} + R^{\mathcal{M}_\tau}}. \quad (3)$$

3.2 Metrics as Utility Functions for Translation Re-Ranking

Let us consider the set $T = \{t_1, t_2, ..., t_n\}$ containing translations of the same source text. We are interested in assessing metric performance in ranking the best translation, as determined by human annotators, in the first position. However, metrics and humans might return tied assessments, placing two or more translations together in the first position. Therefore, we define $T^{\mathcal{M}}$ as the subset of T containing all translations assigned with the highest score by $\mathcal{M}$. Similarly, $T^{\mathcal{H}}$ contains the translations of T ranked highest by human annotators. The Re-Ranking Precision of metric $\mathcal{M}$ is defined as follows:

$$RRP^{\mathcal{M}} = \frac{|T^{\mathcal{M}} \cap T^{\mathcal{H}}|}{|T^{\mathcal{M}}|}. \quad (4)$$

Unlike in the data filtering scenario, we focus solely on Re-Ranking Precision, not Recall. This is because, to serve as a proxy for translation re-ranking applications, what matters is whether the returned translation is the best – or among the best – rather than identifying all the translations ranked highest by human annotators.

4 Experimental Setup

This section outlines the data employed, the metrics evaluated, and our methodology. Implementation details regarding the calculation of Precision, Recall, and F-score are in Appendix A.

4.1 The Data

We employ **WMT23$_{\text{MQM}}$** (Freitag et al., 2023), which contains human annotations collected within the Multidimensional Quality Metrics framework (Lommel et al., 2014, MQM), and **WMT23$_{\text{DA+SQM}}$** (Kocmi et al., 2023), which includes human annotations as Direct Assessments + Scalar Quality Metrics (Kocmi et al., 2022, DA+SQM). Both datasets consist of source texts translated by multiple MT systems, with translation quality assessed by professional human annotators. Table 4 in the Appendix provides additional information regarding these datasets.

We conduct the evaluation using WMT23$_{\text{MQM}}$, specifically the ZH→EN language direction, and use WMT23$_{\text{DA+SQM}}$ as a reference for human performance, given that it contains a subset of the translations in WMT23$_{\text{MQM}}$ annotated using a different human evaluation scheme. Results concerning the other language directions in WMT23$_{\text{MQM}}$, i.e., EN→DE and HE→EN, are reported in the Appendix.

4.2 The Metrics

We consider the following metrics: COMET, COMET-QE, and COMET-QE-MQM (Rei et al., 2020, 2021); BLEURT-20 (Sellam et al., 2020; Pu et al., 2021); MetricX-23, MetricX-23-QE, MetricX-23-XL, and MetricX-23-QE-XL (Juraska et al., 2023); CometKiwi and CometKiwi-XL (Rei et al., 2022, 2023a); XCOMET-ENSEMBLE and XCOMET-QE-ENSEMBLE (Guerreiro et al., 2024); XCOMET-XL (Guerreiro et al., 2024); MaTESe and MaTESe-QE (Perrella et al., 2022); GEMBA-MQM (Kocmi and Federmann, 2023); MBR-MetricX-QE (Naskar et al., 2023). We refer the reader to Appendix C for detailed information regarding these metrics, where we also provide a broader selection, including lexical-based and sentinel metrics (Perrella et al., 2024).

Additionally, following Freitag et al. (2023), we include the results from a random baseline, i.e., Random-sysname, which outputs discrete scores drawn from several Gaussian distributions, one for each MT system that translated the texts in the test set. Each Gaussian has a randomly assigned mean between 0 and 9, with a standard deviation of 2.

4.3 Selecting the Thresholds τ

In the data filtering scenario, we can measure two different aspects of metric performance, depending on how we select the τ value:

1. By selecting τ to maximize the F-score **on the test set**, we measure MT metrics' ability to separate GOOD from BAD translations under ideal conditions. This scenario allows us to measure the maximum achievable F-score for each metric on the test data, effectively evaluating the metric's discriminative power. Metrics whose assessments are not accurate enough, noisy, or, more generally, poorly aligned with human judgments, will achieve a lower optimal F-score than the others.

2. By selecting τ to maximize the F-score **on a development set**, we estimate the true values of MT metrics' Precision, Recall, and F-score in data filtering applications.

We measure metric performance in both evaluation scenarios. As a development set, we use **WMT22$_{\text{MQM}}$**, which contains MQM-based human annotations (Freitag et al., 2022b). However, since some of the tested metrics were trained using WMT22$_{\text{MQM}}$ data, we restrict this experiment to the other metrics.

4.4 Extracting Binary Labels from Manually-Annotated Datasets

Within the MQM annotation framework, professional annotators identify span-level translation errors and assign each error a category and severity. The final MQM score is calculated based on these errors using the following weighting scheme (Freitag et al., 2021a):

Error severity	Category	Penalty
Major	Non-translation	-25
	Others	-5
Minor	Punctuation	-0.1
	Others	-1

We map the annotations in WMT23$_{\text{MQM}}$ and WMT22$_{\text{MQM}}$ to binary labels by considering translations with a score above a certain threshold as GOOD. Specifically, if a translation is assigned an MQM score h, we label it as GOOD if $h \geq -4$, meaning it contains no Major errors and at most 4 Minor ones (or more if Minor punctuation errors are present). Additionally, we classify translations as PERFECT if they contain at most 1 Minor error, i.e., those with $h \geq -1$.[4] This allows us to investigate metrics' ability to distinguish between PERFECT and OTHER translations.

5 Results

In this section, we report the performance obtained by MT metrics when used as binary classifiers to distinguish between GOOD and BAD, as well as PERFECT and OTHER translations, and in terms of their effectiveness in translation re-ranking, i.e., in selecting the best translations among candidates for the same source text.

[4]We use $h \geq -1$ and not $h = 0$ because the inter-annotator agreement in MT evaluation is not particularly high (Freitag et al., 2021a), even with high-cost annotation frameworks like MQM. Therefore, we argue that selecting only translations with a score of 0 might overly depend on individual annotators' preferences.

5.1 Binary Classification

Table 1 shows MT metrics' threshold values, Precision, Recall, and F-score in distinguishing GOOD from BAD and PERFECT from OTHER translations, with the optimal threshold τ selected on the test set. As can be seen, most MT metrics perform reasonably well in distinguishing between GOOD and BAD translations, achieving optimal F-scores as high as 81.59 and 81.40, from GEMBA-MQM and XCOMET-QE-ENSEMBLE, respectively, and as low as 75.81, from BLEURT-20. Instead, lower performance is observed when differentiating between PERFECT and OTHER translations, with the highest F-score being 68.47, from XCOMET-ENSEMBLE. We also note that Precision is almost always lower than Recall, despite the optimal threshold τ being selected to maximize F_β-score with $\beta = \frac{1}{\sqrt{2}}$, which gives more weight to Precision over Recall. These results suggest that the metrics may lack the sensitivity required to distinguish between high-quality translations that differ in minor nuances rather than major errors. As a result, they may resort to lower thresholds, compensating for their lack of Precision with a higher Recall.

Table 2 reports threshold values, Precision, Recall, and F-score when the threshold is optimized on the development set. Note that we restrict the set of tested metrics to those that are openly available and do not employ the WMT22$_{\text{MQM}}$ data for training. As expected, the F-score values are lower than the optimal ones reported in Table 1. Nonetheless, the metric rankings remain stable across the two settings, with MetricX-23-XL and MetricX-23-QE-XL outperforming the other metrics among the reference-based and reference-free ones, respectively.

In general, it is worth noting that the best-performing openly available, reference-free metric is MetricX-23-QE-XL. This result is consistent across language pairs (Appendix D). Therefore, we recommend using MetricX-23-QE-XL for data filtering applications.

Thresholds reliability Our results show that optimal thresholds tend to vary when moving from WMT23$_{\text{MQM}}$ to WMT22$_{\text{MQM}}$ for their calculation. For example, MetricX-23-QE-XL's τ shifts from -3.57 to -5.45, and from -1.64 to -3.54, when separating between GOOD and BAD and PERFECT and OTHER, respectively, and other metrics display a similar pattern. Optimal threshold

	Metric	GOOD vs BAD				PERFECT vs OTHER				Re-ranking	
		τ	P	R	F	τ	P	R	F	RRP	Avg.
REFERENCE BASED	XCOMET-ENSEMBLE	0.83	79.91	84.42	81.36	0.91	68.25	68.93	68.47	43.17	−2.38
	XCOMET-XL	0.80	78.33	83.63	80.02	0.92	67.55	67.46	67.52	37.49	−2.75
	MetricX-23	−4.79	77.43	86.23	80.15	−2.25	63.99	73.20	66.79	39.63	−2.72
	MetricX-23-XL	−3.52	77.80	84.46	79.90	−1.74	65.60	72.54	67.76	39.52	−2.71
	MaTESe	−4.00	76.53	78.10	77.05	−1.00	55.75	79.88	61.99	33.07	−3.18
	COMET	0.76	74.56	78.76	75.91	0.82	61.25	64.38	62.26	34.25	−3.06
	BLEURT-20	0.60	72.76	82.76	75.81	0.67	55.88	69.21	59.71	33.35	−3.07
REFERENCE FREE	XCOMET-QE-ENSEMBLE	0.83	80.40	83.47	81.40	0.92	70.00	63.60	67.73	41.40	−2.47
	MBR-MetricX-QE	0.73	79.00	82.81	80.23	0.80	67.02	65.91	66.64	38.47	−2.40
	MetricX-23-QE	−3.90	76.73	87.70	80.07	−1.31	67.76	67.85	67.79	37.55	−2.59
	MetricX-23-QE-XL	−3.57	77.91	83.36	79.64	−1.64	67.15	70.08	68.10	36.09	−2.83
	GEMBA-MQM	−5.00	82.41	79.99	81.59	−1.00	64.12	74.12	67.14	42.58	−2.30
	MaTESe-QE	−4.00	73.72	85.64	77.30	0.00	55.43	75.05	60.72	30.34	−3.59
	COMET-QE	−0.01	75.35	82.53	77.60	0.05	59.64	68.59	62.35	37.35	−2.66
	COMET-QE-MQM	0.08	75.40	86.33	78.72	0.10	61.63	73.84	65.22	33.52	−3.59
	CometKiwi	0.76	78.62	80.90	79.37	0.80	64.79	66.52	65.35	39.28	−2.61
	CometKiwi-XL	0.64	78.04	79.81	78.62	0.71	64.73	65.51	64.99	38.78	−2.60
	Random-sysname	−5.00	64.06	100.00	72.78	−4.00	42.14	99.99	52.21	29.04	−3.74
	DA+SQM	63.50	67.83	95.95	75.18	74.67	48.30	82.61	56.06	32.99	−3.22

Table 1: Metrics' Precision, Recall, and F-score in binary classification, distinguishing GOOD from BAD, and PERFECT from OTHER translations. τ is selected to maximize the F-score **on the test set**. In the last two columns, we report metrics' Precision in translation re-ranking and the average MQM score of the selected translations. The test set is $\text{WMT23}_{\text{MQM}}$ and the translation direction is ZH→EN. We report results concerning other translation directions in Appendix D. The metrics highlighted in grey are not openly available.[5]

values do not appear stable across language pairs either, as illustrated in Figures 3 and 4 in the Appendix. Specifically, optimal thresholds frequently differ between EN→DE and the other translation directions in the GOOD vs BAD scenario, and are substantially lower for HE→EN in the PERFECT vs OTHER scenario. Such differences in the optimal thresholds might suggest that metric scores have different meanings depending on the translation direction. Furthermore, and as already discussed, an overly small optimal threshold might suggest that metrics are not precise enough. However, threshold values might also be influenced by the quality of the translations in the dataset. Indeed, on average, the HE→EN dataset contains higher-quality translations compared to the other language pairs (Table 5 in the Appendix). This might incentivize the metrics to "settle" for lower threshold values in order to maximize Recall.

In general, we believe that the characteristics of the development set play an important role in determining appropriate thresholds for data filtering applications. Aside from the translation direction, evaluation datasets can differ in the MT systems employed to translate the source texts, the data domains included, and the human annotators who assessed translation quality. Therefore, while leaving the investigation of these phenomena to future work, we recommend estimating optimal thresholds using as much annotated data as possible, to prevent the peculiarities of any single dataset from overly influencing the estimated values. To support this, we are releasing our evaluation framework with options for estimating optimal metric thresholds across several datasets, depending on the user's Precision and Recall requirements.

5.1.1 What is the human performance?

As a reference for human performance, we examine the agreement between two annotation schemas:

[5]Due to its dependence on GPT-4, we include GEMBA-MQM in the group of not openly available metrics. Instead, openly-available versions of MetricX-23 and MetricX-23-QE can be found at https://github.com/google-research/metricx. However, we could not compute their predictions due to their high parameter count and thus resorted to using their outputs as submitted to WMT23, which were computed using different checkpoints than those currently available. Therefore, we include MetricX-23 and MetricX-23-QE in the group of not openly available metrics as well.

	Metric	GOOD vs BAD				PERFECT vs OTHER			
		τ	P	R	F	τ	P	R	F
REFERENCE BASED	MetricX-23-XL	−3.93	76.58	87.01	79.77	−2.97	57.70	88.80	65.32
	COMET	0.77	75.95	75.28	75.72	0.79	55.51	74.57	60.68
	BLEURT-20	0.61	73.81	79.92	75.74	0.64	52.45	76.89	58.67
REFERENCE FREE	MetricX-23-QE-XL	−5.45	73.36	91.88	78.64	−3.54	55.63	90.32	63.80
	COMET-QE-MQM	0.07	72.57	92.41	78.17	0.08	52.59	93.19	61.53
	COMET-QE	−0.02	73.77	85.81	77.39	−0.02	50.54	88.44	58.96
	CometKiwi	0.74	75.57	85.35	78.58	0.76	53.38	86.73	61.23
	CometKiwi-XL	0.62	75.47	83.37	77.93	0.64	53.70	85.03	61.22

Table 2: Metrics' Precision, Recall, and F-score in binary classification, distinguish GOOD from BAD, and PERFECT from OTHER translations. τ is selected to maximize the F-score **on the development set**. The test set is $\text{WMT23}_{\text{MQM}}$ and the translation direction is ZH→EN. Results in other translation directions are in Appendix D.

DA+SQM and MQM. We use DA+SQM as a metric and show the results in the last row of Table 1. It is important to note that DA+SQM annotations do not fully encompass the set of MQM annotations. Consequently, we evaluate its performance on a subset of the data, including $\approx 70\%$ of the data used for metrics, which means that their scores are not directly comparable. However, we observe that DA+SQM performs poorly in absolute terms, with particularly low Precision and a much higher Recall. We hypothesize that this might be due to DA+SQM's annotation guidelines, which instruct annotators to assign only a general quality score to translations rather than identifying specific translation errors, as is done within the MQM framework.[6] This might lead to noisy annotations, rendering DA+SQM less suitable for fine-grained translation quality assessments.

We further investigate this in Appendix F by restricting the data available to MT metrics to that of DA+SQM, to estimate their segment-level correlations with MQM scores. We find DA+SQM annotations correlate less strongly with MQM compared to all tested automatic metrics. We wish to emphasize that Kocmi et al. (2023) have already noted that DA+SQM-based annotations exhibit reduced precision in distinguishing between MT systems with similar performance, compared to MQM-based annotations. Furthermore, in a recent study, Kocmi et al. (2024b) compared the correlation of several human annotation schemes with MQM and found that DA+SQM performed poorly. In line with these findings, our results raise additional concerns regarding DA+SQM reliability, showing performance inferior to automatic metrics.

5.1.2 How BAD are false positives?

Our analysis suggests that MT metrics struggle to achieve high precision in binary classification. Concerning this, we are interested in assessing how *bad* the false positives are – i.e., translations that metrics mislabel as GOOD or PERFECT. To this end, we plot in Figure 2 the distributions of the MQM score Δ computed for a false positive as the difference between the MQM score and the human threshold, which is -4 for a GOOD translation and -1 for a PERFECT one.

The average false positive Δ ranges from -4.25 to -2.85 for both GOOD and PERFECT classifications, indicating that the translations mislabeled by the best metrics contain an average of ≈ 3 additional Minor errors. Overall, the MQM Δ distributions of top-performing metrics are low-variance and skewed to the right, particularly when classifying PERFECT translations. In contrast, less accurate metrics exhibit high-variance distributions, with the average Δ shifting towards lower values. Again, DA+SQM performance is notably poor, showing the highest-variance distribution and the most leftward-shifted average.

5.2 Translation Re-Ranking

We present the results in the last two columns of Table 1. To facilitate interpretation of these results, we also calculate the average MQM score of the translations ranked highest by MT metrics, and report it in the last column. Additionally, it is important to note that there are 15 translations per

[6] Within DA+SQM, human annotators rate a translation from 0 to 100 using a slider with 7 marked levels, where each level is paired with a description of the corresponding translation quality.

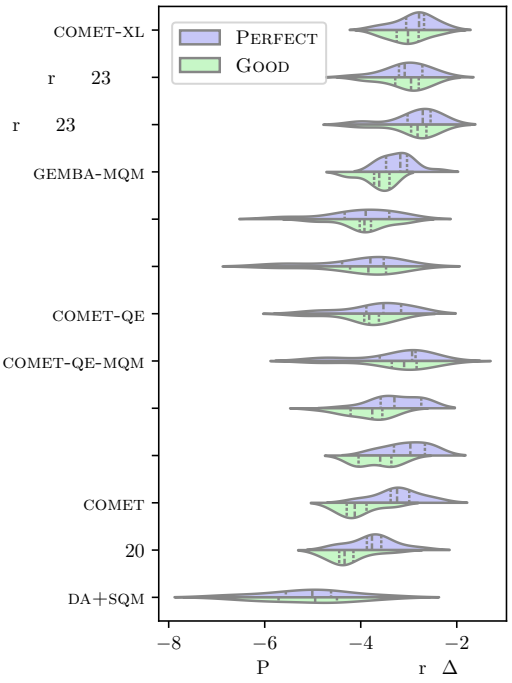

Figure 2: Distribution of the MQM score Δ between the openly available metrics' false positive MQM scores and human thresholds, i.e., -4 for GOOD and -1 for PERFECT. The dataset employed is the ZH→EN split of $\text{WMT23}_{\text{MQM}}$. Additional metrics are included in Figure 5 in the Appendix.

source text.

As shown, metric precision ranges from 30% to 43%, with the highest precision rates being 43.17% and 42.58%, achieved by XCOMET-ENSEMBLE and GEMBA-MQM, respectively. Furthermore, the top-performing metrics yield average MQM scores of ≈ -2.50, indicating that their highest-ranked translations contain an average of two and a half Minor errors. In contrast, human judgments suggest that the average MQM score of the highest-ranked translations is -0.67.

Notably, reference-based metrics consistently outperform reference-free ones. By looking at pairs of metrics of the same family, XCOMET-ENSEMBLE, MetricX-23, and MetricX-23-XL outperform XCOMET-QE-ENSEMBLE, MetricX-23-QE, and MetricX-23-QE-XL, respectively, in ZH→EN and across the other translation directions (see Tables 7 and 8 for results concerning EN→DE and HE→EN). However, in real-world translation re-ranking scenarios, references are not available. To account for this, we assess the performance of reference-based metrics when used as the utility function in an MBR decoding-like scenario, therefore not relying on the presence of reference translations.[7]

	ZH→EN		EN→DE		HE→EN	
Metric	**MBR**	**Tab 1**	**MBR**	**Tab 7**	**MBR**	**Tab 8**
XCOMET-XL	40.27	37.49	44.03	47.31	65.05	68.31
MetricX-23-XL	41.10	39.52	48.24	47.81	63.71	67.17
MaTESe	35.27	33.07	44.53	43.18	60.20	61.99
COMET	37.20	34.25	45.12	48.26	66.38	70.01
BLEURT-20	36.07	33.35	47.76	48.27	63.70	68.33
#1 REF FREE	–	39.28	–	45.71	–	65.61
#2 REF FREE	–	38.78	–	45.57	–	63.25

Table 3: Re-Ranking Precision of reference-based metrics when used as the utility function for MBR decoding, compared with the reference-based re-ranking scenario in the last two columns of Tables 1, 7, and 8. The last two rows of this table show the performance of the best and second-best reference-free metrics in translation re-ranking.

Metric performance in MBR decoding Table 3 shows the translation re-ranking performance of reference-based metrics when used as the utility function for MBR decoding, and comparing it to the reference-based re-ranking scenario presented in the last columns of Tables 1, 7, and 8. On average, the absence of reference translations reduces performance. However, this is not true for all language directions, as MBR decoding outperforms reference-based re-ranking in ZH→EN. We believe this exception is due to the particularly poor quality of the references in the ZH→EN split of $\text{WMT23}_{\text{MQM}}$, as discussed by Freitag et al. (2023).

Furthermore, the last two rows of Table 3 present the performance of the best and second-best openly available, reference-free metrics in translation re-ranking. Our results indicate that reference-based metrics used as the utility function for MBR decoding tend to outperform reference-free metrics. Specifically, the best performance is achieved by two reference-based metrics: MetricX-23-XL, in ZH→EN and EN→DE, and COMET in HE→EN. Again, by looking at metrics of the same family, MetricX-23-XL outperforms MetricX-23-QE-XL across the board. These findings suggest that translation re-ranking using MBR decoding may be more reliable than QE re-ranking.[8] Previous stud-

[7] MBR decoding seeks the candidate translation that maximizes an external notion of utility (Eikema and Aziz, 2022). The set of candidate translations is used as both the set of hypotheses and to approximate a set of references. Then, each candidate translation is compared with all the others, employing reference-based metrics as the utility function.

[8] This is based on the assumption that using translations generated by distinct MT systems can serve as an effective ap-

ies have already compared MT metrics in MBR decoding and QE re-ranking. For example, Freitag et al. (2022a) and Fernandes et al. (2022) employ human annotators to evaluate the quality of translations produced by MT systems when using one or the other re-ranking technique. However, due to the high cost of human annotations, these studies were limited to including only a few metrics. Furthermore, their findings may need to be revisited to determine whether they remain valid with the introduction of new metrics. In contrast, by relying solely on the human annotations released annually at WMT, our setup facilitates updating results as soon as new metrics or datasets become available.

6 Related Work

Previous studies have focused primarily on the problem of *Error attribution*. Specifically, the Shared Task on Quality Estimation at WMT investigated the ability of MT metrics to predict word-level annotations (Zerva et al., 2022; Blain et al., 2023). Fomicheva et al. (2022a) and Rei et al. (2023b) employed attribution methods to derive explanations for the predictions of MT metrics, measuring the faithfulness of such explanations by comparing them to human annotations. To tackle the same issue, Perrella et al. (2022), Fernandes et al. (2023), Guerreiro et al. (2024), Kocmi and Federmann (2023), and Xu et al. (2023) proposed metrics that address the lack of *Error attribution* by providing explanations in the form of either span-level annotations or natural language rationales. Furthermore, recent studies have introduced dedicated benchmarks to investigate the impact of specific translation errors, such as disambiguation errors (Campolungo et al., 2022; Martelli et al., 2024) and wrongly translated named entities (Conia et al., 2024).

In a different vein, and closer to our work, some studies explored the meaning of raw metrics scores in terms of their alignment with human judgments. Mathur et al. (2020a) studied the meaning of system-level score deltas for BLEU (Papineni et al., 2002), showing that a statistically significant increase of 0-3 BLEU points corresponds to significantly better MT systems less than half of the time, in terms of human judgments. Similarly, Kocmi et al. (2024a) investigated the relationship between MT metrics' system-level score deltas and human judgments. Finally, in a recent study, Agrawal et al. (2024) evaluated MT metrics' ability to assess high-quality translations by examining their correlation with human judgments, as well as their Precision, Recall, and F-score, using a setup similar to ours. However, instead of calculating metrics thresholds from data, they arbitrarily assumed that a metric indicates *high-quality* only if its normalized assessments fall within the $[0.99, 1.00]$ interval. In contrast, we measure metrics performance in data filtering without making assumptions about the meaning of their assessments, aiming to understand this meaning through the evaluation itself.

7 Conclusion

In this work, we introduce a novel evaluation framework for MT metrics. Within this framework, we measure metrics performance in i) binary classification, i.e., distinguishing between GOOD and BAD, and PERFECT and OTHER translations, and ii) in a proxy scenario for translation re-ranking, selecting the best among the translations of the same source text. By measuring performance in terms of Precision, Recall, and F-score, we fulfill a dual purpose. First, we offer a more intuitive interpretation of metrics' capabilities, as compared to correlation with human judgment, and second, we provide concrete user recommendations concerning novel MT metric use cases. We find that MT metrics perform relatively well in distinguishing between GOOD and BAD translations, but struggle with Precision, especially when dealing with higher-quality translations like in the PERFECT vs OTHER scenario. Our results show that MetricX-23-QE-XL is the best openly available metric for data filtering applications, while MetricX-23-XL and COMET achieve the highest performance in translation re-ranking. Additionally, we demonstrate that reference-based MT metrics, when used as the utility function in an MBR decoding-like scenario, outperform reference-free ones, suggesting that MBR decoding may be superior to QE re-ranking. Finally, we report notably poor performance for DA+SQM annotations used as a metric within our evaluation framework, raising concerns about its reliability.

proximation of sampling from a single system. We delve into the differences between MBR decoding and our evaluation setup in the Limitations.

8 Limitations

Language coverage We acknowledge that the scope of our work is limited by the available test data, covering only a few language directions. However, our evaluation framework is agnostic to the test data employed. Therefore, we leave the investigation of metric performance in more language directions to future works, depending on the availability of new annotated datasets.

Design choices in the data filtering scenario We made certain arbitrary decisions in the design of our framework and experimental setup. We chose F_β-score to select the optimal threshold τ, with $\beta = \frac{1}{\sqrt{2}}$. While we explained our reasons for giving Precision a higher weight than Recall, it remains unclear whether $\beta = \frac{1}{\sqrt{2}}$ is the optimal choice. Furthermore, we selected the human score thresholds to be -4, for GOOD translations, and -1 for PERFECT ones. We recognize that practitioners might have different requirements and may want to narrow or broaden these definitions. Therefore, we release our evaluation framework leaving this as an option for users.

Evaluation fairness in the data filtering scenario In one of the two setups proposed, we selected the threshold τ to maximize the F-score on the test set used for the evaluation. This optimization process might favor metrics whose assessments are more sensitive to the underlying gold score distribution, enabling them to achieve a better balance between Precision and Recall. As a result, discrete metrics – i.e., those that output scores within a discrete set, such as the integers in $[-25, 0]$ for GEMBA-MQM– might be disadvantaged compared to continuous metrics – i.e., those that output scores within a continuous interval, such as the real values in $[0, 1]$ for metrics of the COMET family. However, we argue that this limitation is inherent to the nature of discrete metrics rather than a flaw in our evaluation framework. Indeed, studying the ability of MT metrics to distinguish between GOOD and BAD translations requires identifying the score threshold that best separates them, and discrete metrics inherently offer a much more limited set of options for optimizing this threshold. Nonetheless, if discrete metrics are indeed disadvantaged, using a development set could mitigate the impact of this phenomenon.

Alignment between the translation re-ranking scenario and the corresponding metric use cases We designed the translation re-ranking scenario as a proxy for QE re-ranking and MBR decoding. However, our setup differs from these two use cases in two ways:

1. Candidates number: The test datasets we used feature 15, 12, and 13 translations per source text, for ZH→EN, EN→DE, and HE→EN, respectively. However, in QE re-ranking and MBR decoding it is common to work with a larger number of candidate translations, often reaching hundreds per source text.

2. Candidates selection: In QE re-ranking and MBR decoding, candidate translations are typically sampled from the same MT system. In contrast, in our annotated datasets, each candidate translation was generated by a different MT system.

In future work, it would be interesting to investigate whether our results might vary when dealing with a higher number of candidate translations or when all candidates are sampled from the same MT system.

Acknowledgements

We gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR, and the CREATIVE project (CRoss-modal understanding and gEnerATIon of Visual and tExtual content), which is funded by the MUR Progetti di Rilevante Interesse Nazionale programme (PRIN 2020).

This work was carried out while Lorenzo Proietti was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome.

References

Sweta Agrawal, António Farinhas, Ricardo Rei, and André F. T. Martins. 2024. Can automatic metrics assess high-quality translations? *arXiv preprint, arXiv:2405.18348*.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual

large language model for translation-related tasks. In *Proceedings of the First Conference on Language Modeling*.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Berlin, Germany. Association for Computational Linguistics.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *Proceedings of the International Conference on Learning Representations*.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929, Singapore. Association for Computational Linguistics.

Sören Dreano, Derek Molloy, and Noel Murphy. 2023. Tokengram_F, a fast and accurate token-based chrF++ derivative. In *Proceedings of the Eighth Conference on Machine Translation*, pages 730–737, Singapore. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2022. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Muhammad ElNokrashy and Tom Kocmi. 2023. eBLEU: Unexpectedly good machine translation evaluation using simple word embeddings. In *Proceedings of the Eighth Conference on Machine Translation*, pages 746–750, Singapore. Association for Computational Linguistics.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. Association for Computational Linguistics.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth*

Conference on Machine Translation, pages 1066–1083, Singapore. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Mara Finkelstein and Markus Freitag. 2024. MBR and QE finetuning: Training-time distillation of the best and most expensive decoding methods. In *Proceedings of The Twelfth International Conference on Learning Representations*.

Marina Fomicheva, Lucia Specia, and Nikolaos Aletras. 2022a. Translation error detection as rationale extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4148–4159, Dublin, Ireland. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022b. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

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

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022a. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022b. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.

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.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced self-training (rest) for language modeling. *arXiv preprint, arXiv:2308.08998*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of The Eleventh International Conference on Learning Representations*.

Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a

reward model. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8164–8180, Mexico City, Mexico. 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.

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). 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.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024a. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. *arXiv preprint, arXiv:2406.11580*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Belgium, Brussels. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Federico Martelli, Stefano Perrella, Niccolò Campolungo, Tina Munda, Svetla Koeva, Carole Tiberius, and Roberto Navigli. 2024. DiBiMT: A Gold Evaluation Benchmark for Studying Lexical Ambiguity in Machine Translation. *Computational Linguistics*, pages 1–79.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020a. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020b. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Subhajit Naskar, Daniel Deutsch, and Markus Freitag. 2023. Quality estimation using minimum Bayes risk. In *Proceedings of the Eighth Conference on Machine Translation*, pages 806–811, Singapore. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. Technical report, OpenAI.

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.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, Bangkok, Thailand. Association for Computational Linguistics.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022. MaTESe: Machine translation evaluation as a sequence tagging problem. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 569–577, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, and Markus Freitag. 2023. There's no data like better data: Using QE metrics for MT data filtering. In *Proceedings of the Eighth Conference on Machine Translation*, pages 561–577, Singapore. 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.

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Miguel Ramos, Patrick Fernandes, António Farinhas, and Andre Martins. 2024. Aligning neural machine translation models: Human feedback in training and inference. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 258–274, Sheffield, UK. European Association for Machine Translation (EAMT).

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, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023a. 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, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. 2023b. The inside story: Towards better understanding of machine translation neural evaluation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105, Toronto, Canada. 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. 2022. 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.

Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the WMT15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint, arXiv:2207.04672*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint, arXiv:2401.08417*.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Singapore. 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.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proceedings of the International Conference on Learning Representations*.

A Implementation Details

Most annotated datasets used for metric evaluation – such as $\text{WMT23}_{\text{MQM}}$ and $\text{WMT22}_{\text{MQM}}$– contain a selection of source texts translated by multiple MT systems. As a result, each source text is paired with several automatic translations, along with one or more manually-curated references. In this respect, and to align the data filtering scenario to its real use case,[9] we group the translations according to the MT system (or human annotator) that generated them, computing Precision and Recall on each group. In MT meta-evaluation, this strategy is called *Group-by-System* or *System Grouping*, by Deutsch et al. (2023) and Perrella et al. (2024), respectively. Finally, we aggregate these statistics across systems obtaining average Precision and Recall measures, which are used to obtain the F-score as in Equation 3.

Instead, concerning the translation re-ranking scenario, translations are grouped according to their source text. This strategy is called *Group-by-Item* or *Segment Grouping* by Deutsch et al. (2023) and Perrella et al. (2024), respectively. Consequently, the final Re-Ranking Precision is the average across the Precision values computed for each source text as in Equation 4.

Selecting the Thresholds τ For each metric, we select the threshold τ that maximizes the F-score, either on the test or development set. To find the optimal threshold for a metric, we i) collect all its assessments on the considered dataset, removing duplicates; ii) measure Precision, Recall, and F-score corresponding to each candidate threshold value; and iii) select the threshold that yields the highest F-score.

B Datasets Statistics

Table 4 presents the number of systems, segments, and annotations in $\text{WMT23}_{\text{MQM}}$ and $\text{WMT23}_{\text{DA+SQM}}$.

Table 5 presents the average and median MQM scores assigned to the translations in $\text{WMT23}_{\text{MQM}}$ and $\text{WMT22}_{\text{MQM}}$.

C The Metrics

We consider the following metrics:

- COMET, COMET-QE, and COMET-QE-MQM, (Rei et al., 2020, 2021) are a reference-based and two reference-free regression-based metrics, respectively, built upon the XLM-RoBERTa large architecture (Conneau et al., 2020), and trained using datasets containing human annotations in the form of Direct Assessments (DA) (Graham et al., 2013). Specifically, COMET was trained on the datasets released at WMT between 2017 and 2020 (Bojar et al., 2017; Ma et al., 2018, 2019; Mathur et al., 2020b), while COMET-QE and COMET-QE-MQM also include the DA-based datasets released in 2015 and 2016 (Stanojević et al., 2015; Bojar et al., 2016). COMET-QE-MQM was further fine-tuned on a split of the MQM-based corpus released by Freitag et al. (2021a).[10]

- BLEURT-20 (Sellam et al., 2020; Pu et al., 2021) is a regression-based metric built upon the RemBERT pre-trained language model

[9] When used for data filtering, MT metrics filter parallel corpora that typically contain only one translation per source text.

[10] https://github.com/Unbabel/COMET. We used the version 2.2.1 of the COMET framework.

	ZH→EN			EN→DE			HE→EN		
	Sys.	**Seg.**	**Total**	**Sys.**	**Seg.**	**Total**	**Sys.**	**Seg.**	**Total**
$\text{WMT23}_{\text{MQM}}$	15	1177	17655	12	460	5520	13	820	10660
$\text{WMT23}_{\text{DA+SQM}}$	15	884	13260	12	460	5520	–	–	–

Table 4: Number of MT systems, source segments, and the total number of annotations in $\text{WMT23}_{\text{MQM}}$ and $\text{WMT23}_{\text{DA+SQM}}$, excluding official WMT23 references employed by reference-based metrics. Concerning $\text{WMT23}_{\text{DA+SQM}}$, we restricted the annotations to those available in $\text{WMT23}_{\text{MQM}}$ and discarded the rest.

	ZH→EN		EN→DE		HE→EN	
	Avg.	**Median**	**Avg.**	**Median**	**Avg.**	**Median**
$\text{WMT23}_{\text{MQM}}$	−4.21	−2.00	−7.47	−3.00	−2.35	0.00
$\text{WMT22}_{\text{MQM}}$	−3.21	−1.00	−1.31	0.00	–	–

Table 5: Average and Median MQM scores of the translations in $\text{WMT23}_{\text{MQM}}$ and $\text{WMT22}_{\text{MQM}}$.

(Chung et al., 2021). RemBERT was fine-tuned on DA-based human assessments from 2015 to 2019, along with synthetic data.[11]

- BERTscore (Zhang et al., 2020) leverages pre-trained encoders to extract the contextualized embeddings of the tokens of a translation and its reference. Then, it computes the cosine similarity between each pair of embeddings, greedily matching the most similar ones.[12]

- MetricX-23, MetricX-23-QE, MetricX-23-XL and MetricX-23-QE-XL (Juraska et al., 2023) are regression-based metrics built upon the mT5-XXL (the first two) and mT5-XL (the last two) models (Xue et al., 2021). These metrics are trained using DA-based human judgments released at WMT between 2015 and 2020 (Stanojević et al., 2015; Bojar et al., 2016), and further fine-tuned on a combination of MQM-based human judgments and synthetic data (Freitag et al., 2021a,b).[13]

- CometKiwi and CometKiwi-XL (Rei et al., 2022, 2023a) are reference-free regression-based metrics, built upon InfoXLM (Chi et al., 2021) and XLM-RoBERTa XL (Goyal et al., 2021), respectively. These metrics are trained using DA-based human judgments released at WMT from 2017 to 2020, as well as DA from the MLQE-PE corpus (Fomicheva et al., 2022b). CometKiwi-XL's training data also include the DA for Indian languages released by Blain et al. (2023). [14]

- XCOMET-XL (Guerreiro et al., 2024) is a regression-based metric built upon the XLM-RoBERTa XL architecture, trained on the concatenation of DA-based human judgments released at WMT from 2017 and 2020 and the MLQE-PE dataset, and further fine-tuned using MQM-based annotations coming from the following datasets: i) WMT data from 2020 to 2022, ii) IndicMT (Sai B et al., 2023), and iii) DEMETR (Karpinska et al., 2022). Given a candidate translation, XCOMET-XL jointly identifies its error spans and assigns it a scalar quality score. XCOMET-ENSEMBLE and XCOMET-QE-ENSEMBLE are ensembles between one XL and two XXL XCOMET checkpoints that result from different training stages.[15]

- MaTESe and MaTESe-QE (Perrella et al., 2022) are a reference-based and a reference-free metric, respectively, built upon InfoXLM and DeBERTaV3 (He et al., 2023). MaTESe metrics annotate the spans of translations that contain an error, specifying the error severity.[16]

- GEMBA-MQM (Kocmi and Federmann, 2023) is an LLM-based metric that leverages GPT-4 to return quality assessments in the form of

[11] https://github.com/google-research/bleurt.
[12] https://github.com/Tiiiger/bert_score.
[13] https://github.com/google-research/metricx.
[14] See footnote 10.
[15] See footnote 10.
[16] https://github.com/SapienzaNLP/MaTESe.

MQM annotations.[17]

- MBR-MetricX-QE (Naskar et al., 2023) is based on the MBR decoding strategy. Given a translation, it uses an MT system to generate pseudo-references and a reference-based MT metric (MetricX-23) as the MBR utility function.

- BLEU (Papineni et al., 2002) is a precision-oriented metric that computes the number of overlapping n-grams between a translation and its reference.[18]

- chrF (Popović, 2015) compares a translation and its reference based on the number of overlapping character n-grams.[19]

- f200spBLEU (Goyal et al., 2022; Team et al., 2022) computes BLEU scores using subword tokenization done by the standardized FLORES-200 Sentencepiece models.[20]

- eBLEU (ElNokrashy and Kocmi, 2023) matches the n-grams of semantically similar words between a candidate translation and a reference using non-contextual word embeddings.[21]

- tokengram_F (Dreano et al., 2023) is derived from chrF++ (Popović, 2017) by replacing word-based n-grams with token-based n-grams, as obtained from popular tokenization algorithms such as BPE (Sennrich et al., 2016) or Unigram (Kudo, 2018).[22]

In addition, we include three sentinel metrics, i.e., metrics designed explicitly to detect issues with the meta-evaluation (Perrella et al., 2024):

- SENTINEL$_{\text{CAND}}$ assesses the quality of a translation without taking its source or reference as input.

- SENTINEL$_{\text{SRC}}$ predicts the quality of a translation based solely on its source, without taking the translation itself as input.

- SENTINEL$_{\text{REF}}$ predicts the quality of a translation based solely on its reference, without taking the translation itself as input.

Trained with incomplete information, these metrics are not supposed to rank high in a fair meta-evaluation setup. Sentinel metrics are regression-based and were trained using WMT data. In particular, they were trained using DA annotations from 2017 to 2020 and further fine-tuned with MQM scores from 2020 to 2022.

D Additional Results

In this section, we report all our results considering all the language directions available in WMT23$_{\text{MQM}}$, i.e., ZH→EN, EN→DE, and HE→EN, and including all MT metrics mentioned in Appendix C.

Tables 6, 7, and 8 show the performance of MT metrics in the data filtering scenario when τ is selected as the one that maximizes F-score on the test set, i.e., WMT23$_{\text{MQM}}$. The last two columns contain the performance of MT metrics in the translation re-ranking scenario. Instead, Tables 9 and 10 show the performance of MT metrics in the data filtering scenario when τ is selected as the one that maximizes the F-score on the development set, i.e., WMT22$_{\text{MQM}}$, and the performance is measured on WMT23$_{\text{MQM}}$.

The performance of lexical-based metrics All lexical-based metrics fail, partially or completely, at tackling the data filtering task. In most cases, their optimal threshold is 0.0[23], indicating that they lack the sensitivity required to separate GOOD from BAD and PERFECT from OTHER translations, and therefore resort to maximizing recall. Instead, lexical-based metrics achieve a decent performance in the translation re-ranking scenario. Nonetheless, they still perform worse than most neural-based metrics.

The performance of sentinel metrics As illustrated in Appendix A, we use the *System Grouping* strategy to align the data filtering scenario to its real use case. Specifically, we compute Precision and Recall on the translations of each MT system independently and then compute final statistics by averaging them across MT systems. As demonstrated by Perrella et al. (2024), this setting is particularly susceptible to spurious correlations in the

[17] https://github.com/MicrosoftTranslator/GEMBA.
[18] https://github.com/mjpost/sacrebleu.
[19] See footnote 18.
[20] See footnote 18.
[21] https://github.com/munael/ebleu-mt-metrics-wmt23.
[22] https://github.com/SorenDreano/tokengram_F.
[23] Note that these metrics' score range is either $[0, 1]$ or $[0, 100]$.

evaluation data, which favor trained metrics over the rest. As a consequence, the performance of sentinel metrics is not as low as it would be in a fair evaluation scenario. However, we highlight that this scenario was intentionally designed to adhere closely to the data filtering use case, and adopting a different grouping strategy could reduce its effectiveness as a proxy for this task. Therefore, we argue that careful attention should be given to selecting source texts in evaluation datasets, with the goal of minimizing the impact of spurious correlations and ultimately ensuring that sentinel metrics rank at the bottom of the metric rankings. Nonetheless, despite sentinel metrics performing better than they ideally ought to, they still do not surpass most state-of-the-art metrics, differently from the results obtained by Perrella et al. (2024). Similarly, GEMBA-MQM performs decently in many of our settings, whereas Perrella et al. (2024) report it ranking lower than sentinel metrics when using *System Grouping* (specifically, in two out of three translation directions, namely ZH→EN and EN→DE).[24] Given these observations, we believe that the binary classification setup lessens the impact of spurious correlations, as compared to the correlation with human judgment.

Instead, and as expected, sentinel metrics rank at the bottom in translation re-ranking. Indeed, the translation re-ranking scenario involves selecting the best among the translations of the same source text, i.e., using the *Segment Grouping* strategy, which, as shown by Perrella et al. (2024), counters the impact of spurious correlations in the evaluation dataset.

E Additional Figures

In Figures 3 and 4, we report metrics optimal threshold values across different language directions. The thresholds were selected to maximize the F-score on the test set.

In Figure 5 we report the Δ MQM score between the false positives and the human thresholds in the GOOD and PERFECT translations classification scenario.

[24] Since GEMBA-MQM was not fine-tuned using human assessments, it should not be able to leverage spurious correlations in metrics' training data to conduct the evaluation. Perrella et al. (2024) report GEMBA-MQM ranking lower than sentinel metrics, suggesting that the evaluation might unfairly favor metrics that have learned spurious correlations during training.

F DA+SQM and MQM Correlation

Tables 11 and 12 present the segment-level correlation between the tested metrics and MQM when considering DA+SQM as a metric. We employ Pearson's ρ and Kendall's τ correlation coefficients, and acc_{eq} accuracy (Deutsch et al., 2023). As recommended by Perrella et al. (2024), we use *Segment Grouping*, meaning that we compute these statistics on groups of translations of the same source text, and then average them. To enable a fair comparison between metrics and DA+SQM, we restrict the evaluation datasets to the translations with available DA+SQM annotations.

	Metric	GOOD vs BAD				PERFECT vs OTHER				Re-ranking	
		τ	P	R	F	τ	P	R	F	RRP	Avg.
REFERENCE BASED	xCOMET-ENSEMBLE	0.83	79.91	84.42	81.36	0.91	68.25	68.93	68.47	43.17	−2.38
	xCOMET-XL	0.80	78.33	83.63	80.02	0.92	67.55	67.46	67.52	37.49	−2.75
	MetricX-23	−4.79	77.43	86.23	80.15	−2.25	63.99	73.20	66.79	39.63	−2.72
	MetricX-23-XL	−3.52	77.80	84.46	79.90	−1.74	65.60	72.54	67.76	39.52	−2.71
	MaTESe	−4.00	76.53	78.10	77.05	−1.00	55.75	79.88	61.99	33.07	−3.18
	COMET	0.76	74.56	78.76	75.91	0.82	61.25	64.38	62.26	34.25	−3.06
	BLEURT-20	0.60	72.76	82.76	75.81	0.67	55.88	69.21	59.71	33.35	−3.07
	BERTscore	0.84	64.33	99.47	72.91	0.92	48.20	69.15	53.62	32.29	−3.20
REFERENCE FREE	xCOMET-QE-ENSEMBLE	0.83	80.40	83.47	81.40	0.92	70.00	63.60	67.73	41.40	−2.47
	MBR-MetricX-QE	0.73	79.00	82.81	80.23	0.80	67.02	65.91	66.64	38.47	−2.40
	MetricX-23-QE	−3.90	76.73	87.70	80.07	−1.31	67.76	67.85	67.79	37.55	−2.59
	MetricX-23-QE-XL	−3.57	77.91	83.36	79.64	−1.64	67.15	70.08	68.10	36.09	−2.83
	GEMBA-MQM	−5.00	82.41	79.99	81.59	−1.00	64.12	74.12	67.14	42.58	−2.30
	MaTESe-QE	−4.00	73.72	85.64	77.30	0.00	55.43	75.05	60.72	30.34	−3.59
	COMET-QE	−0.01	75.35	82.53	77.60	0.05	59.64	68.59	62.35	37.35	−2.66
	COMET-QE-MQM	0.08	75.40	86.33	78.72	0.10	61.63	73.84	65.22	33.52	−3.59
	CometKiwi	0.76	78.62	80.90	79.37	0.80	64.79	66.52	65.35	39.28	−2.61
	CometKiwi-XL	0.64	78.04	79.81	78.62	0.71	64.73	65.51	64.99	38.78	−2.60
LEXICAL BASED	BLEU	0.00	64.06	100.00	72.78	0.00	42.13	100.00	52.20	30.09	−3.50
	chrF	0.00	64.06	100.00	72.78	0.00	42.13	100.00	52.20	31.51	−3.39
	eBLEU	0.02	64.11	99.82	72.79	0.03	42.20	99.87	52.26	30.16	−3.49
	f200spBLEU	0.00	64.06	100.00	72.78	0.00	42.13	100.00	52.20	30.80	−3.46
	tokengram_F	0.00	64.06	100.00	72.78	0.00	42.13	100.00	52.20	30.55	−3.44
SENTINEL METRICS	$\text{SENTINEL}_{\text{SRC}}$	−0.14	75.64	83.31	78.03	0.23	63.00	71.90	65.71	25.77	−4.21
	$\text{SENTINEL}_{\text{REF}}$	−0.55	71.74	91.25	77.24	0.08	59.11	73.33	63.19	25.77	−4.21
	$\text{SENTINEL}_{\text{CAND}}$	−0.14	75.43	86.92	78.91	0.22	63.16	71.84	65.81	29.38	−3.83
	Random-sysname	−5.00	64.06	100.00	72.78	−4.00	42.14	99.99	52.21	29.04	−3.74
	DA+SQM	63.50	67.83	95.95	75.18	74.67	48.30	82.61	56.06	32.99	−3.22

Table 6: Metrics' Precision, Recall, and F-score in binary classification, distinguishing GOOD from BAD, and PERFECT from OTHER translations. τ is selected to maximize the F-score **on the test set**. In the last two columns, we report metrics' Precision in translation re-ranking and the average MQM score of the selected translations. The test set is $\text{WMT23}_{\text{MQM}}$ and the translation direction is ZH→EN. The metrics highlighted in grey are not openly available.

	Metric	GOOD vs BAD				PERFECT vs OTHER				Re-ranking	
		τ	P	R	F	τ	P	R	F	RRP	Avg.
REFERENCE BASED	xCOMET-ENSEMBLE	0.90	80.55	70.52	76.90	0.92	75.90	67.86	73.02	48.79	−3.58
	xCOMET-XL	0.90	78.49	71.17	75.89	0.94	78.17	65.44	73.41	47.31	−3.91
	MetricX-23	−1.71	79.56	67.61	75.14	−1.18	75.21	68.18	72.71	52.61	−3.47
	MetricX-23-XL	−1.55	77.62	73.98	76.37	−1.07	72.59	70.88	72.01	47.81	−3.74
	MaTESe	−1.00	71.87	72.92	72.21	0.00	67.42	60.67	65.01	43.18	−4.56
	COMET	0.83	72.56	70.81	71.97	0.88	81.10	55.66	70.38	48.26	−3.50
	BLEURT-20	0.70	76.49	69.54	74.03	0.74	72.66	63.26	69.23	48.27	−3.64
	BERTscore	0.85	57.66	81.22	63.83	0.92	68.12	40.42	55.45	43.11	−4.55
REFERENCE FREE	xCOMET-QE-ENSEMBLE	0.87	79.99	70.39	76.51	0.91	75.58	66.10	72.13	46.70	−3.90
	MBR-MetricX-QE	0.76	78.82	70.50	75.84	0.80	74.25	66.75	71.57	48.81	−3.78
	MetricX-23-QE	−2.07	75.65	75.84	75.71	−1.09	76.81	64.88	72.37	48.04	−3.58
	MetricX-23-QE-XL	−1.99	75.86	73.18	74.94	−1.27	74.08	68.44	72.10	45.57	−3.96
	GEMBA-MQM	−1.00	79.69	66.77	74.86	0.00	75.07	62.04	70.16	42.52	−4.04
	MaTESe-QE	−2.00	67.48	82.89	71.93	0.00	68.16	63.48	66.52	41.03	−5.14
	COMET-QE	0.04	68.75	73.61	70.30	0.07	65.26	62.99	64.49	45.71	−3.84
	COMET-QE-MQM	0.08	74.18	73.98	74.12	0.09	73.50	65.64	70.68	41.25	−4.82
	CometKiwi	0.82	75.86	68.51	73.24	0.82	64.08	71.37	66.34	41.75	−4.32
	CometKiwi-XL	0.69	73.52	70.71	72.56	0.73	67.44	64.13	66.30	43.67	−4.45
LEXICAL BASED	BLEU	3.29	52.19	99.25	61.99	0.00	39.95	100.00	49.94	42.95	−4.28
	chrF	28.67	52.59	99.47	62.39	73.11	62.32	39.61	52.32	41.43	−4.48
	eBLEU	0.14	52.12	99.63	61.97	0.75	62.15	37.38	50.90	40.86	−4.76
	f200spBLEU	7.02	52.56	98.56	62.25	48.17	53.42	51.43	52.74	43.69	−4.21
	tokengram_F	0.29	52.60	99.18	62.36	0.69	54.63	47.20	51.91	42.63	−4.43
SENTINEL METRICS	$\text{SENTINEL}_{\text{SRC}}$	0.22	75.96	69.88	73.82	0.37	73.80	62.93	69.78	30.98	−7.47
	$\text{SENTINEL}_{\text{REF}}$	0.17	73.12	68.86	71.65	0.27	68.76	66.90	68.13	30.98	−7.47
	$\text{SENTINEL}_{\text{CAND}}$	0.20	75.73	68.80	73.27	0.28	68.70	69.00	68.80	43.93	−4.72
	Random-sysname	−4.00	52.07	99.94	61.96	−4.00	39.96	99.92	49.95	40.56	−5.36
	DA+SQM	77.33	59.60	85.39	66.27	82.67	48.90	77.21	55.71	37.11	−5.01

Table 7: Metrics' Precision, Recall, and F-score in binary classification, distinguishing GOOD from BAD, and PERFECT from OTHER translations. τ is selected to maximize the F-score **on the test set**. In the last two columns, we report metrics' Precision in translation re-ranking and the average MQM score of the selected translations. The test set is $\text{WMT23}_{\text{MQM}}$ and the translation direction is EN→DE. The metrics highlighted in grey are not openly available.

	Metric	GOOD vs BAD				PERFECT vs OTHER				Re-ranking	
		τ	P	R	F	τ	P	R	F	RRP	Avg.
REFERENCE BASED	xCOMET-ENSEMBLE	0.84	83.28	85.81	84.10	0.87	83.42	80.69	82.49	69.21	−0.99
	xCOMET-XL	0.81	82.00	87.14	83.64	0.85	81.24	82.82	81.76	68.31	−0.98
	MetricX-23	−4.26	82.09	86.51	83.51	−3.34	81.78	81.96	81.84	67.20	−1.11
	MetricX-23-XL	−3.44	81.94	87.23	83.63	−3.39	79.11	88.20	81.92	67.17	−1.07
	MaTESe	−4.00	84.49	76.57	81.67	−4.00	81.86	77.99	80.53	61.99	−1.51
	COMET	0.77	78.96	87.93	81.74	0.81	78.95	81.31	79.72	70.01	−1.03
	BLEURT-20	0.67	80.85	83.39	81.68	0.67	77.48	84.02	79.55	68.33	−1.04
	BERTscore	0.92	76.33	88.14	79.90	0.93	73.89	85.97	77.52	69.88	−0.88
REFERENCE FREE	xCOMET-QE-ENSEMBLE	0.82	80.92	86.90	82.82	0.85	80.96	80.60	80.84	66.22	−1.40
	MBR-MetricX-QE	0.74	81.57	86.17	83.05	0.74	78.01	86.65	80.69	68.09	−1.25
	MetricX-23-QE	−1.79	80.27	89.66	83.17	−1.28	79.42	85.60	81.38	63.17	−1.46
	MetricX-23-QE-XL	−3.46	80.32	86.03	82.13	−3.19	78.27	85.08	80.41	63.25	−1.64
	GEMBA-MQM	−7.00	79.70	89.24	82.64	−5.00	79.01	83.83	80.55	65.22	−1.26
	MaTESe-QE	−6.00	74.35	95.01	80.16	−3.00	75.99	81.43	77.72	53.95	−2.28
	COMET-QE	−0.03	75.62	91.71	80.32	−0.00	74.21	86.43	77.88	61.06	−1.70
	COMET-QE-MQM	0.08	76.28	89.22	80.16	0.09	74.16	87.36	78.09	52.57	−2.32
	CometKiwi	0.77	80.14	86.48	82.14	0.80	80.02	79.85	79.96	60.42	−1.55
	CometKiwi-XL	0.60	77.83	89.83	81.46	0.63	76.90	84.57	79.30	65.61	−1.28
LEXICAL BASED	BLEU	0.00	71.31	100.00	78.85	0.00	67.89	100.00	76.03	64.46	−1.38
	chrF	0.00	71.31	100.00	78.85	0.00	67.89	100.00	76.03	65.61	−1.27
	eBLEU	0.02	71.36	99.94	78.88	0.02	67.93	99.94	76.05	65.29	−1.32
	f200spBLEU	0.00	71.31	100.00	78.85	7.35	68.71	96.65	76.04	66.04	−1.23
	tokengram_F	0.00	71.31	100.00	78.85	0.00	67.89	100.00	76.03	65.30	−1.24
SENTINEL METRICS	$\text{SENTINEL}_{\text{SRC}}$	−0.10	74.72	91.11	79.48	−0.01	72.89	88.48	77.44	53.09	−2.35
	$\text{SENTINEL}_{\text{REF}}$	−0.78	73.28	96.26	79.62	−0.78	70.04	96.66	77.12	53.09	−2.35
	$\text{SENTINEL}_{\text{CAND}}$	−0.52	74.56	93.53	79.97	−0.50	71.42	93.81	77.59	45.30	−3.04
	Random-sysname	−5.00	71.32	99.99	78.85	−5.00	67.89	99.99	76.03	53.51	−2.08

Table 8: Metrics' Precision, Recall, and F-score in binary classification, distinguishing GOOD from BAD, and PERFECT from OTHER translations. τ is selected to maximize the F-score **on the test set**. In the last two columns, we report metrics' Precision in translation re-ranking and the average MQM score of the selected translations. The test set is $\text{WMT23}_{\text{MQM}}$ and the translation direction is HE→EN. The metrics highlighted in grey are not openly available.

	Metric	GOOD vs BAD				PERFECT vs OTHER			
		τ	P	R	F	τ	P	R	F
REFERENCE BASED	MetricX-23-XL	−3.93	76.58	87.01	79.77	−2.97	57.70	88.80	65.32
	COMET	0.77	75.95	75.28	75.72	0.79	55.51	74.57	60.68
	BLEURT-20	0.61	73.81	79.92	75.74	0.64	52.45	76.89	58.67
	BERTscore	0.92	71.44	63.39	68.54	0.93	49.99	59.81	52.88
REFERENCE FREE	MetricX-23-QE-XL	−5.45	73.36	91.88	78.64	−3.54	55.63	90.32	63.80
	COMET-QE-MQM	0.07	72.57	92.41	78.17	0.08	52.59	93.19	61.53
	COMET-QE	−0.02	73.77	85.81	77.39	−0.02	50.54	88.44	58.96
	CometKiwi	0.74	75.57	85.35	78.58	0.76	53.38	86.73	61.23
	CometKiwi-XL	0.62	75.47	83.37	77.93	0.64	53.70	85.03	61.22
LEXICAL BASED	f200spBLEU	4.86	64.52	89.01	71.03	4.86	42.53	89.35	51.53
	BLEU	5.44	64.78	85.99	70.58	5.43	42.73	86.42	51.39
	chrF	2.08	64.04	99.84	72.73	2.08	42.09	99.77	52.14

Table 9: Metrics' Precision, Recall, and F-score in binary classification, distinguish GOOD from BAD, and PERFECT from OTHER translations. τ is selected to maximize the F-score **on the development set**, i.e., $\text{WMT22}_{\text{MQM}}$. The test set is $\text{WMT23}_{\text{MQM}}$ and the translation direction is ZH→EN.

	Metric	GOOD vs BAD				PERFECT vs OTHER			
		τ	P	R	F	τ	P	R	F
REFERENCE BASED	MetricX-23-XL	−2.10	73.10	81.34	75.65	−1.10	71.84	71.41	71.70
	COMET	0.72	57.94	94.05	66.44	0.80	53.88	84.32	61.25
	BLEURT-20	0.60	63.64	87.97	70.11	0.66	56.99	82.09	63.45
	BERTscore	0.68	52.15	99.84	62.03	0.68	40.00	99.79	49.99
REFERENCE FREE	MetricX-23-QE-XL	−2.71	70.68	81.67	74.00	−1.66	67.12	76.24	69.91
	COMET-QE-MQM	0.07	72.32	76.27	73.59	0.09	69.07	71.16	69.75
	COMET-QE	−0.05	56.09	94.26	64.84	−0.01	47.75	87.66	56.29
	CometKiwi	0.73	60.98	90.91	68.50	0.79	55.90	83.61	62.84
	CometKiwi-XL	0.52	59.13	91.29	67.00	0.62	51.57	86.45	59.59
LEXICAL BASED	f200spBLEU	3.18	52.15	99.74	62.01	3.67	40.05	99.79	50.04
	BLEU	0.00	52.05	100.00	61.95	0.00	39.95	100.00	49.94
	chrF	0.00	52.05	100.00	61.95	0.00	39.95	100.00	49.94

Table 10: Metrics' Precision, Recall, and F-score in binary classification, distinguish GOOD from BAD, and PERFECT from OTHER translations. τ is selected to maximize the F-score **on the development set**, i.e., $\text{WMT22}_{\text{MQM}}$. The test set is $\text{WMT23}_{\text{MQM}}$ and the translation direction is EN→DE.

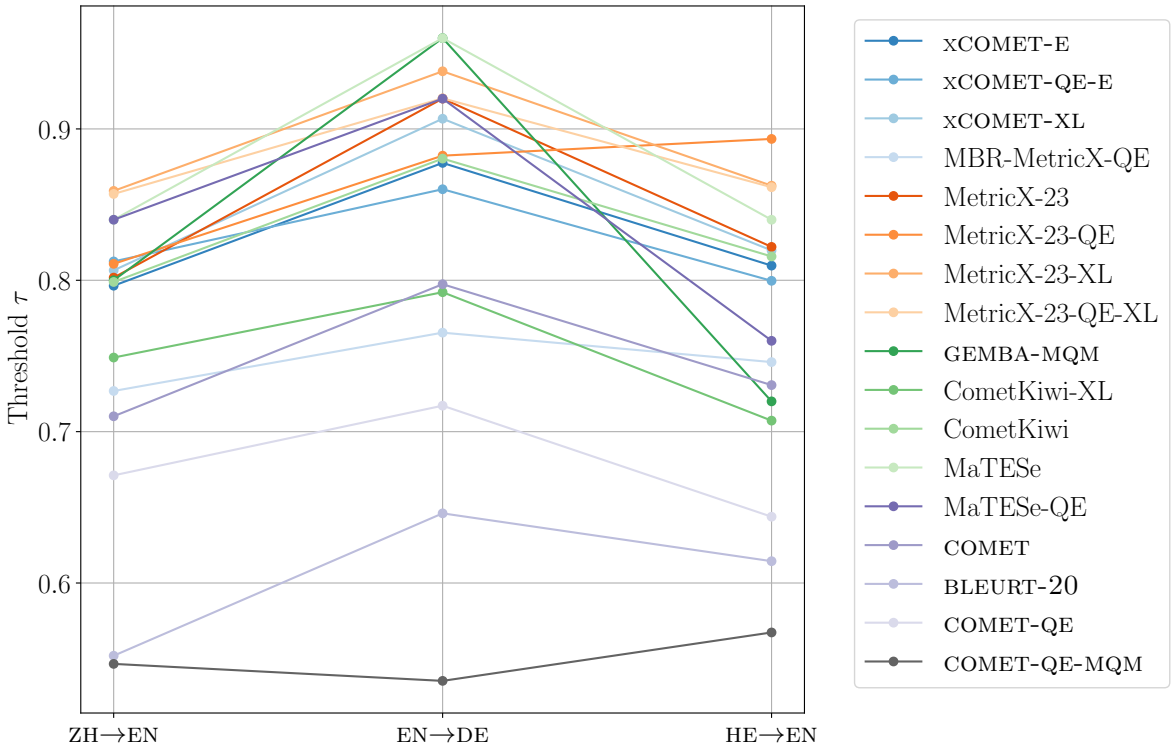

Figure 3: Tested metrics' optimal threshold values across different language directions. The thresholds were selected to maximize the F-score on the test set in the GOOD vs BAD binary classification scenario. Thresholds are normalized between 0 and 1 for improved clarity.

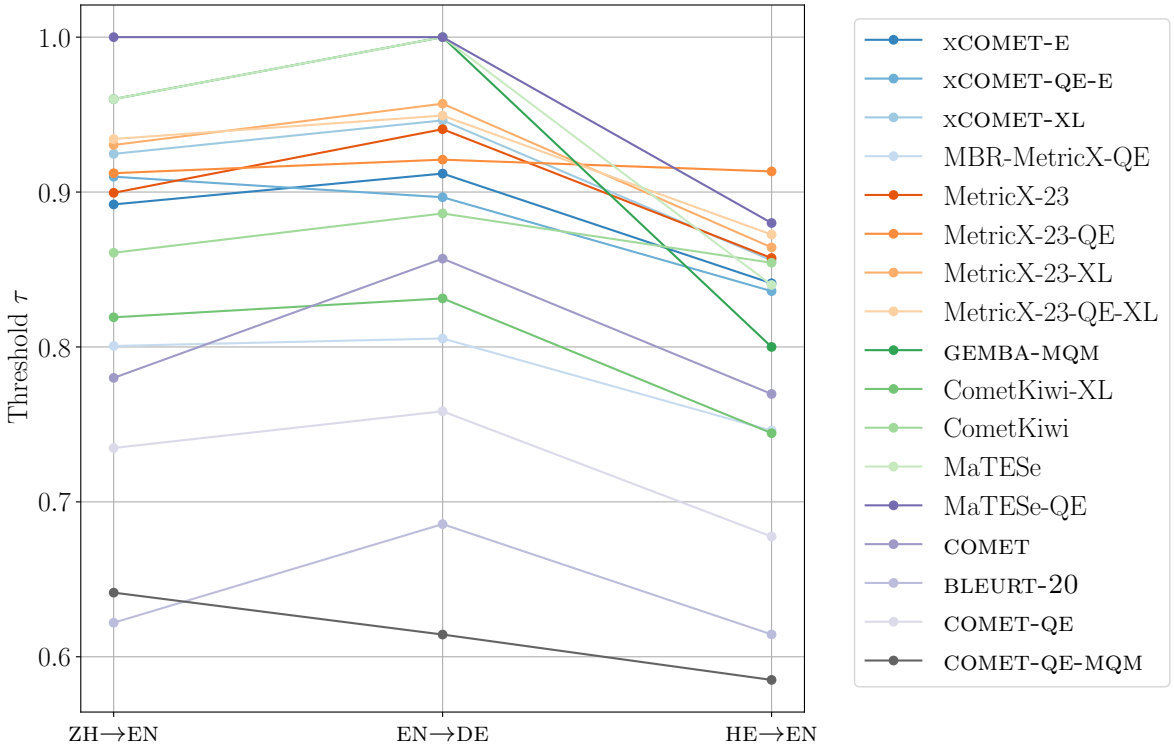

Figure 4: Tested metrics' optimal threshold values across different language directions. The thresholds were selected to maximize the F-score on the test set in the PERFECT vs OTHER binary classification scenario. Thresholds are normalized between 0 and 1 for improved clarity.

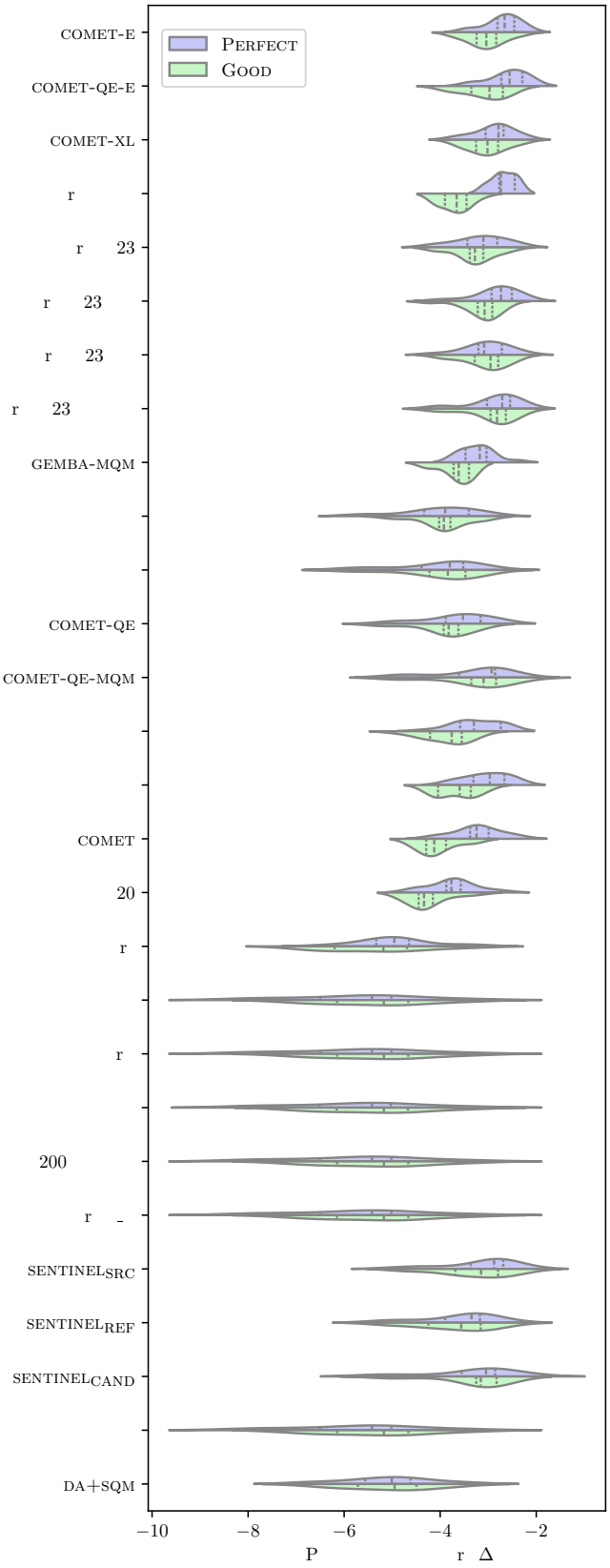

Figure 5: Distribution of the MQM score Δ between metrics' false positive MQM scores and human thresholds, i.e., -4 for GOOD and -1 for PERFECT. The dataset is the ZH→EN split of WMT23$_{MQM}$.

Metric	τ	ρ	acc_{eq}
GEMBA-MQM	0.36	0.43	0.52
xCOMET-ENSEMBLE	0.30	0.42	0.54
xCOMET-QE-ENSEMBLE	0.26	0.37	0.53
xCOMET-XL	0.26	0.38	0.52
MBR-MetricX-QE	0.29	0.43	0.53
MetricX-23	0.26	0.37	0.53
MetricX-23-QE	0.24	0.35	0.52
MetricX-23-XL	0.25	0.36	0.52
CometKiwi	0.25	0.37	0.52
CometKiwi-XL	0.25	0.38	0.52
COMET	0.25	0.36	0.51
BLEURT-20	0.26	0.37	0.52
MaTESe	0.27	0.33	0.48
COMET-QE-MQM	0.16	0.21	0.48
MaTESe-QE	0.21	0.24	0.44
DA+SQM	0.11	0.20	0.42
Random-sysname	0.02	0.02	0.38

Table 11: Kendall τ and Pearson ρ correlation coefficients, and acc_{eq} accuracy (Deutsch et al., 2023), measured between the DA+SQM- and MQM-based annotations, and between MT metrics and MQM. The data is the intersection between $\text{WMT23}_{\text{MQM}}$ and $\text{WMT23}_{\text{DA+SQM}}$. The language direction is ZH→EN.

Metric	τ	ρ	acc_{eq}
GEMBA-MQM	0.40	0.48	0.57
MBR-MetricX-QE	0.40	0.54	0.58
xCOMET-ENSEMBLE	0.38	0.54	0.60
xCOMET-XL	0.37	0.51	0.60
MetricX-23	0.37	0.51	0.60
COMET	0.37	0.51	0.58
BLEURT-20	0.37	0.49	0.57
MetricX-23-XL	0.36	0.49	0.59
xCOMET-QE-ENSEMBLE	0.36	0.51	0.59
MetricX-23-QE	0.36	0.51	0.60
COMET-QE	0.35	0.47	0.57
MetricX-23-QE-XL	0.35	0.45	0.59
CometKiwi-XL	0.35	0.50	0.57
CometKiwi	0.33	0.46	0.57
COMET-QE-MQM	0.29	0.39	0.54
MaTESe	0.29	0.33	0.53
MaTESe-QE	0.28	0.34	0.52
DA+SQM	0.17	0.29	0.46
Random-sysname	0.08	0.12	0.41

Table 12: Kendall τ and Pearson ρ correlation coefficients, and acc_{eq} accuracy (Deutsch et al., 2023), measured between the DA+SQM- and MQM-based annotations, and between MT metrics and MQM. The data is the intersection between $\text{WMT23}_{\text{MQM}}$ and $\text{WMT23}_{\text{DA+SQM}}$. The language direction is EN→DE.