Guardians of the Machine Translation Meta-Evaluation: Sentinel Metrics Fall In!

Stefano Perrella[1,*] **Lorenzo Proietti**[1,*] **Alessandro Scirè**[1,2]
Edoardo Barba[1] **Roberto Navigli**[1]
[1]Sapienza NLP Group, Sapienza University of Rome
[2]Babelscape, Italy
{perrella, lproietti, scire, barba, navigli}@diag.uniroma1.it

Abstract

Annually, at the Conference of Machine Translation (WMT), the Metrics Shared Task organizers conduct the meta-evaluation of Machine Translation (MT) metrics, ranking them according to their correlation with human judgments. Their results guide researchers toward enhancing the next generation of metrics and MT systems. With the recent introduction of neural metrics, the field has witnessed notable advancements. Nevertheless, the inherent opacity of these metrics has posed substantial challenges to the meta-evaluation process. This work highlights two issues with the meta-evaluation framework currently employed in WMT, and assesses their impact on the metrics rankings. To do this, we introduce the concept of sentinel metrics, which are designed explicitly to scrutinize the meta-evaluation process's accuracy, robustness, and fairness. By employing sentinel metrics, we aim to validate our findings, and shed light on and monitor the potential biases or inconsistencies in the rankings. We discover that the present meta-evaluation framework favors two categories of metrics: i) those explicitly trained to mimic human quality assessments, and ii) continuous metrics. Finally, we raise concerns regarding the evaluation capabilities of state-of-the-art metrics, emphasizing that they might be basing their assessments on spurious correlations found in their training data.

1 Introduction

Over the past few years, the Machine Translation (MT) field has witnessed significant advancements, largely driven by the advent of neural architectures, with the Transformer (Vaswani et al., 2017) being the most notable. Modern MT systems deliver mostly fluent and accurate translations, posing a challenge for their quality evaluation – even when conducted by human annotators, especially those who lack professional training (Freitag et al., 2021a). Under these circumstances, shallow overlap-based metrics are gradually being replaced by neural-based metrics, which demonstrate a better correlation with human judgments (Freitag et al., 2022). However, a significant limitation is that most neural-based metrics are black-box systems trained to predict human judgments in the form of scalar scores, and typically do not provide justifications for their assessments. Besides rendering them challenging to interpret, such opacity also complicates their meta-evaluation. In this respect, we found that certain strategies for the assessment of MT metrics' capabilities – which have recently been employed in the context of the Metrics Shared Task at the Conference on Machine Translation (WMT)[1] – favor specific metric categories and potentially encourage undesirable metrics behavior. To demonstrate these problems, we introduce the concept of sentinel metrics, i.e., a suite of metrics serving as a probe to identify pitfalls in the meta-evaluation process. Sentinel metrics are either trained with incomplete information – which makes them inherently unable to evaluate the quality of machine-translated text properly – or consist of variations of existing metrics – which have been devised to expose specific issues in the meta-evaluation.

As an example, in Table 1, we present the segment-level ranking of WMT23 with the inclusion of a sentinel metric. As can be seen, $\text{SENTINEL}_{\text{CAND}}$ ranks in the upper half. $\text{SENTINEL}_{\text{CAND}}$ is a sentinel metric designed to assess the quality of a candidate translation based solely on the translation itself, without accessing its source sentence or any reference translation. Arguably, such a metric should only be capable of evaluating a translation's fluency, but not its ad-

[*]Equal contribution.

[1]With its first edition in 2006 (Koehn and Monz, 2006), "WMT is the main event for machine translation and machine translation research." (https://machinetranslate.org/wmt).

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 16216–16244

Metric		Avg. corr
XCOMET-Ensemble	1	0.697
MetricX-23	2	0.682
XCOMET-QE-Ensemble*	3	0.681
MetricX-23-QE*	4	0.681
mbr-metricx-qe*	5	0.652
GEMBA-MQM*	6	0.639
MaTESe	7	0.636
CometKiwi*	8	0.632
sescoreX	9	0.628
SENTINEL$_{CAND}$*	10	0.626
cometoid22-wmt22*	11	0.625
KG-BERTScore*	12	0.624
COMET	13	0.622
BLEURT-20	14	0.622
Calibri-COMET22-QE*	15	0.603
Calibri-COMET22	16	0.603
YiSi-1	17	0.600
docWMT22CometDA	18	0.598
docWMT22CometKiwiDA*	19	0.598
prismRef	20	0.593
MS-COMET-QE-22*	21	0.588
BERTscore	22	0.582
mre-score-labse-regular	23	0.558
XLsim	24	0.544
f200spBLEU	25	0.540
MEE4	26	0.539
tokengram_F	27	0.537
chrF	28	0.537
BLEU	29	0.533
prismSrc*	30	0.530
embed_llama	31	0.529
eBLEU	32	0.491
Random-sysname*	33	0.463

Table 1: Segment-level ranking of the primary submissions to the WMT 2023 Metrics Shared Task, with the inclusion of sentinel metrics. The values in the column "Avg. corr" are obtained by averaging the correlations of the 6 segment-level tasks of WMT 2023. Starred metrics are reference-free, underlined metrics are baselines, and highlighted metrics are sentinels. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2023), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021). In Table 3 in Appendix A, we report the metrics' performance in terms of rank and correlation in all the 6 tasks that contribute to this ranking. All the rankings present in this work have been computed using the official shared task library (https://github.com/google-research/mt-metrics-eval).

equacy in conveying the original message, and a fair assessment should rank it at lower positions. Notably, SENTINEL$_{CAND}$ is above strong baselines such as COMET (Rei et al., 2020) and BLEURT-20 (Sellam et al., 2020), suggesting that there might be some issues with the segment-level meta-evaluation methods used in WMT23.

In this work, we: i) illustrate the issues that affect the segment-level meta-evaluation measures used in WMT23, demonstrating their impact experimentally with the help of sentinel metrics; ii) propose solutions for addressing these issues; iii) raise concerns regarding the reliability of state-of-the-art MT metrics. We publish the code to reproduce our work and the weights of the sentinel metrics at https://github.com/SapienzaNLP/guardians-mt-eval.

2 The Meta-evaluation of MT Metrics

Yearly, the WMT Metrics Shared Task organizes a competition among metrics, including participants' submissions and baselines, to identify the metric that most closely aligns with human judgments. Historically, the organizers have employed correlation with human judgment as a meta-evaluation strategy. Recently, significant efforts have been made to refine the meta-evaluation process, encompassing the adoption of new measures, such as those proposed by Kocmi et al. (2021) and Deutsch et al. (2023), and the introduction of the challenge sets sub-task (Freitag et al., 2021b, 2022), among other initiatives. In this section, we provide an overview of WMT's official meta-evaluation setting.

First, multiple MT systems are employed to translate source segments found in one or more test datasets.[2] Consequently, test datasets contain several translations of the same source segment. Second, a manual evaluation campaign is carried out to assess the quality of all translations. Finally, metrics' capabilities are assessed based on their alignment with human judgments, which are in the form of scalar scores. Such alignment is typically estimated using correlation and accuracy measures. Specifically, metrics are evaluated at two granularity levels:

- at the segment level, metrics assign a score to every translation, and they are ranked according to their ability to discern between higher- and lower-quality translations;
- at the system level, metrics assign a score to each MT system,[3] and they are ranked according to their ability to discern between superior and inferior systems.

[2]A segment typically refers to a single sentence, but can also include multiple sentences. For instance, at WMT23, a segment represents an entire paragraph rather than a single sentence for the English-to-German translation direction.

[3]Typically, the score of a system is calculated as the mean of the scores given to its translations.

At both granularity levels, metrics can be evaluated using several statistical methods, such as the Kendall τ and Pearson ρ correlation coefficients, which have traditionally been applied at the segment and system levels, respectively. A final metrics ranking is derived by aggregating results from all the chosen statistics. For example, at WMT23, the final ranking was computed from the following three statistics:

1. System-level pairwise ranking accuracy (Kocmi et al., 2021), which evaluates metrics based on their ability to rank systems in the same order as human judgments.

2. System- and segment-level Pearson correlation, which measures the degree to which metric scores and human scores are correlated linearly.

3. Segment-level pairwise ranking accuracy with tie calibration (Deutsch et al., 2023), which evaluates metrics based on their ability to rank segments in the same order as human judgments, or their ability to predict ties correctly.

In this work, we identify two critical issues related to the second and third statistics, and provide the following recommendations to address them:

- **Translations should be grouped by their source segment before calculating segment-level correlations** (Section 3).

- **Tie calibration should not be conducted on the test set** (Section 4).

In the following two sections, we provide an overview of some of the aforementioned statistics, illustrate their flaws, and demonstrate their impact by leveraging our sentinel metrics.

3 To Group or Not to Group?

At early editions of the WMT Metrics Shared Task (Macháček and Bojar, 2013, 2014; Stanojević et al., 2015; Bojar et al., 2016), human assessments were collected in the form of Relative Rankings (RR). Specifically, the annotators were tasked to rank up to 5 translations of the same source sentence, produced by different MT systems. From each ranking, up to 10 pairwise comparisons were extracted. Despite metrics assessments being scalar scores – which theoretically enabled the comparison of all pairs of translated segments – correlation was measured only on those pairs of translations for which RR annotations were available. Therefore, only translations of the same source sentence were compared. Later on, at subsequent editions of WMT, new techniques for human evaluation were adopted: first, Direct Assessments (Graham et al., 2013, DA) – where annotators rate individual translations on a scale from 0 to 100 – then, Multidimensional Quality Metrics (Lommel et al., 2014, MQM) – where annotators tag the spans of a translation that contain errors, specifying their category and severity. With both the new annotation schemas, each translated segment was assigned a scalar quality score independently of the other translations,[4] which made it possible to compare all translations, not only those of the same source sentence. This new possibility raised doubts regarding the best way to compute the correlation between metrics and human assessments. Indeed, it could be computed using all translations at once – *No Grouping* – or by first grouping translations based on either their source segment – *Segment Grouping* – or the system that produced them – *System Grouping* – and then returning the average correlation of these groups.

At the WMT21 Metrics Shared Task, Freitag et al. (2021b) chose the *No Grouping* strategy, arguing that the other options would provide only a partial view of the overall picture. At WMT22, all three grouping strategies were used (Freitag et al., 2022), and later at WMT23, Freitag et al. (2023) chose *No Grouping* again. Although *No Grouping* is the only strategy that assesses the MT metrics' ability to discern between higher- and lower-quality translations in absolute terms, irrespective of the source segment or MT system, we show that both *No Grouping* and *System Grouping* may introduce unfairness and favor trained metrics over the rest.

3.1 The Relation Between Spurious Correlations and Grouping Strategies

Most neural-based metrics are trained with a regression objective to approximate human judgments. They are expected to infer by pattern-matching the relation between human judgments and various phenomena, such as omissions, additions, or other translation errors. However, this mechanism might inadvertently lead to the detection of patterns that are not in a causal relation with the concept of

[4] In MQM, a final score is obtained by applying a specific weighting to each combination of the detected spans' category and severity.

translation quality, but are instead spurious correlations, e.g., the length of a translation, or the number of named entities in it, among others. Arguably, the meta-evaluation should not reward metrics for basing their assessments on spurious correlations between the features of the source, translation, or reference, and the human judgments. However, our intuition is that *No Grouping* and *System Grouping* strategies might be doing so by allowing the comparison of translations from different sources. To simplify, consider a metric that unfairly penalizes a translation solely because it contains many named entities. Using *No Grouping* or *System Grouping*, such a metric might have a non-negative correlation with human judgments if, on average, translating sentences containing many named entities is more challenging than translating other sentences, because MT systems would be making more mistakes in translating them. Therefore, exploiting such a pattern might be beneficial even though it is not causally related to the quality of a translation. In contrast, when using *Segment Grouping*, such a pattern would be ineffective, as different translations of the same source sentence should contain the same amount of named entities. More generally, we would expect *Segment Grouping* to lessen the impact of most spurious correlations derived from features shared by a source sentence and its translations.

To assess the extent of this issue, we incorporate three sentinel metrics into the current meta-evaluation framework and re-compute the metrics' rankings using all grouping strategies. Crucially, we find that the impact of spurious correlations when *No Grouping* and *System Grouping* strategies are employed is substantial – favoring trained metrics over the rest[5] – and is significantly reduced with *Segment Grouping*.

3.2 The Sentinel Metrics

This section describes the three sentinel metrics employed to measure the impact of grouping strategies on the meta-evaluation process:

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

2. SENTINEL$_{\text{SRC}}$, which predicts the quality of a translation solely based on its source.

3. SENTINEL$_{\text{REF}}$, which predicts the quality of a translation solely based on its reference.

Having no information regarding the translation to evaluate, SENTINEL$_{\text{SRC}}$ and SENTINEL$_{\text{REF}}$ can only learn spurious correlations between the features of the source and reference sentences, respectively, and the human judgments. SENTINEL$_{\text{CAND}}$, instead, is a metric with partial information. Indeed, it is possible to evaluate a translation's fluency and grammatical correctness without comparing it with its source or reference sentences, but not its adequacy. Nonetheless, we expect SENTINEL$_{\text{CAND}}$ to base its assessments on spurious correlations also.

3.3 Experimental Setup

Sentinel metrics employ XLM-RoBERTa large (Conneau et al., 2020) as their backbone model, with a multi-layer fully-connected neural network on top of the [CLS] token, which is used to output predictions in the form of scalar scores. We train sentinel metrics to minimize the Mean Squared Error (MSE) between their predicted scores and human judgments. Our dataset comprises a selection of data from WMT spanning 2017 to 2022, incorporating Direct Assessments (DA) and Multidimensional Quality Metrics (MQM) scores. Following Rei et al. (2022a), we train sentinel metrics for a single epoch using DA from 2017 to 2020 and finetune them for a further epoch using MQM data. Additional details regarding the training process are reported in Appendix B.

3.4 Results

In Table 2, we report the ranking derived from the segment-level Pearson correlation of the primary submissions to the Metrics Shared Task of WMT23, with the inclusion of sentinel metrics, in the language direction ZH $\rightarrow$ EN, and with all three grouping strategies. We report in Appendix C the rankings alongside the correlation values for all the official translation directions of the Metrics Shared Task, i.e., ZH $\rightarrow$ EN, EN $\rightarrow$ DE and HE $\rightarrow$ EN. As can be seen, SENTINEL$_{\text{SRC}}$ ranks fourth and third when the grouping strategies are *No Grouping* and *System Grouping*, respectively, surpassing strong baselines like COMET or BLEURT-20, and even state-of-the-art metrics like GEMBA-MQM. The only metrics that are not surpassed are

[5]Indeed, overlap-based metrics such as BLEU (Papineni et al., 2002) and chrF (Popović, 2015), or LLM-based metrics such as GEMBA-MQM (Kocmi and Federmann, 2023), were not trained to mimic human assessments and should not be able to leverage spurious correlations.

large regression-based systems such as XCOMET-Ensemble (Guerreiro et al., 2023) and MetricX-23 (Juraska et al., 2023), which might have learned the same spurious correlations leveraged by the sentinel metrics, in addition to non-spurious patterns (cf. Section 3.4.1). Conversely, when grouping by segment, SENTINEL$_{SRC}$ and SENTINEL$_{REF}$ are correctly positioned at the bottom of the ranking,[6] and SENTINEL$_{CAND}$ ranks 11th, compared to 3rd and 2nd with *No Grouping* and *System Grouping*, respectively. A notable difference between the grouping strategies is the positioning of GEMBA-MQM, which is ranked 7th and 9th with *No Grouping* and *System Grouping*, respectively, and becomes first with *Segment Grouping*. We hypothesize that this is due to GEMBA-MQM being based on GPT-4, which has not been explicitly fine-tuned on human assessments and is less likely to leverage spurious correlations such as those described in Section 3.1. Interestingly, with grouping strategies other than *Segment Grouping*, GEMBA-MQM is surpassed by all the sentinel metrics.

SENTINEL$_{CAND}$ is the only sentinel metric that does not rank at the very bottom with *Segment Grouping*, outperforming prismSrc (Thompson and Post, 2020) and embed_llama (Dreano et al., 2023), and positioning itself within the same cluster of statistical significance as BLEU. This suggests that focusing solely on the candidate translation – specifically, its fluency and grammatical correctness – may be sufficient to exceed the performance of some less effective metrics, at least in terms of Pearson correlation with human judgments. Furthermore, we highlight that our results may provide an answer to the open question left at WMT23 regarding the inconsistency of segment-level and system-level correlations for prismSrc. Freitag et al. (2023) noticed that, despite displaying a moderate correlation at the segment level, prismSrc was showing negative correlation values at the system level. As can be seen from Table 2, prismSrc ranks 15th out of 24 with *No Grouping* but 13th out of 14 with *Segment Grouping* (i.e., it is in the second to last significance cluster, close to the sentinel metrics). This result is consistent with prismSrc's negative correlation at the system level.

In Appendix C, we also report the rankings and correlations obtained using the Kendall τ correlation coefficient for each grouping strategy, to show that our findings are independent of the correlation measure, at least among those typically employed at WMT, i.e., Pearson ρ and Kendall τ.

[6]This had to be expected, given that both these metrics return the same assessment for all translations of the same source segment.

Metric	Grouping No	Seg	Sys
XCOMET-Ensemble	1	2	1
MetricX-23-QE*	1	4	1
XCOMET-QE-Ensemble*	1	3	1
MetricX-23	2	3	2
SENTINEL$_{CAND}$*	3	11	2
SENTINEL$_{SRC}$*	4	14	3
sescoreX	4	7	5
MaTESe	5	6	6
SENTINEL$_{REF}$	5	14	4
mbr-metricx-qe*	6	1	7
cometoid22-wmt22*	6	4	6
GEMBA-MQM*	7	1	9
Calibri-COMET22-QE*	7	5	8
CometKiwi*	7	3	9
KG-BERTScore*	8	4	10
COMET	9	4	12
Calibri-COMET22	9	7	11
docWMT22CometKiwiDA*	10	6	13
BLEURT-20	10	4	13
MS-COMET-QE-22*	11	7	14
docWMT22CometDA	12	6	15
YiSi-1	13	6	16
BERTscore	14	7	17
prismSrc*	15	13	16
prismRef	16	6	18
embed_llama	17	12	18
mre-score-labse-regular	18	8	19
BLEU	19	11	20
XLsim	19	10	21
f200spBLEU	20	10	21
MEE4	20	9	21
chrF	21	8	22
tokengram_F	22	8	23
Random-sysname*	23	14	23
eBLEU	24	10	24

Table 2: Rankings obtained from the segment-level Pearson correlation for the primary submissions to the WMT 2023 Metrics Shared Task, with sentinel metrics. The language direction is ZH → EN. Ranks represent clusters of statistical significance. Additional information can be found in Appendix C.

3.4.1 Are MT metrics learning from spurious correlations?

We hypothesize that some of the trained metrics may be basing their assessments on the same spurious correlations as those leveraged by the sentinel metrics. To delve deeper into this, we measure their segment-level Pearson correlation with the sentinel metrics using *No Grouping*. Surprisingly, XCOMET-Ensemble, XCOMET-QE-Ensemble, MetricX-23, and MetricX-23-QE, which are the only metrics that surpass the sentinels in Table 2, display a high correlation with

all three sentinel metrics. Interestingly, their correlation with $\text{SENTINEL}_{\text{SRC}}$ is 0.750, 0.736, 0.690, and 0.712 (Figure 4), respectively, while their correlation with human judgment is 0.650, 0.647, 0.625, and 0.647, respectively (Table 5). We recognize that these metrics share many similarities with our sentinels, as both are neural transformer-based systems and both were trained with the same regression-based objective, using largely the same data. This similarity likely contributes to the high correlation values observed. However, with access limited to only the source segment, $\text{SENTINEL}_{\text{SRC}}$ relies exclusively on spurious correlations to conduct the evaluation. For this reason, we argue that these results raise concerns about the reliability of state-of-the-art MT metrics, which may be learning to exploit spurious correlations to minimize the Mean Squared Error with human judgments during training. To further support our hypothesis, we plot in Figure 1 the relation between the assessments of XCOMET-Ensemble and translation length, which serves as a simple spurious correlate of translation quality.[7] We also plot the distribution of MQM human judgments over translation length. As we can see from the figure, XCOMET-Ensemble scores decrease at increasing candidate lengths, with the metric almost never assigning scores higher than 0.9 to translations longer than 400 characters. However, the distribution of human judgments shows that human annotators rated many of those translations as perfect or near-perfect, indicating that XCOMET-Ensemble might be biased to assign lower scores to longer translations, irrespective of their quality. Furthermore, the least-squares regression lines show that, on average, and as expected, longer translations contain more errors than shorter ones, and therefore are assigned lower scores by human annotators. This suggests that detecting biases of this type might be particularly complex without datasets crafted specifically for it.

We leave the investigation of these phenomena to future work and, for further details, we direct readers to Appendix D, where we report the pairwise correlation between most of the considered metrics and sentinel metrics, and to appendix E, where we report the relation between such metrics' assessments and translation length.

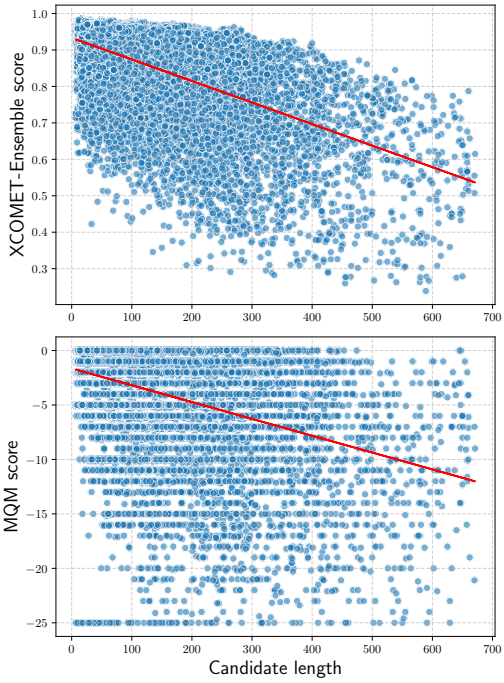


Figure 1: We show XCOMET-Ensemble assessments and MQM-based human judgments in the top and bottom figures, respectively, over the length of the candidate translation (in characters). The red line represents the linear least-squares regression. MQM human judgments smaller than -25 have been removed for improved clarity. The language pair is ZH $\rightarrow$ EN.

4 The Evaluation of Ties

In this Section, we focus on the third statistic among those described in Section 2, i.e., the segment-level pairwise ranking accuracy with tie calibration, dubbed acc_{eq} by Deutsch et al. (2023). Prior to WMT23, the organizers of the Metrics Shared Task used to employ the Kendall τ coefficient – which is a statistic used to estimate the rank-based agreement between two sets of measurements (Kendall, 1945) – to measure the correlation between metrics and human judgments at the segment level. Deutsch et al. (2023) pointed out that the Kendall τ coefficient does not account for metrics correctly predicting ties,[8] and introduced acc_{eq} to address this issue. Unfortunately, our analysis indicates that acc_{eq} inadvertently compromises evaluation fairness in order to accommodate ties, ultimately biasing the results in favor of continuous metrics[9] over discrete ones.

[7] We expect that longer sentences are, on average, more challenging to translate. Therefore, we anticipate that MT metrics might have learned to assign lower scores to longer translations, despite the length and quality of translations not being causally related.

[8] Given a pair of translations whose quality has been assessed by human annotators, the pair is tied if both translations were assigned with the same score.

[9] By continuous, we refer to those metrics whose assessments can take on any value within a given range, as opposed

4.1 The Kendall τ

In this section, we define the Kendall τ coefficient as employed by the organizers of the Metrics Shared Task of WMT21 and WMT22.[10] Let $\boldsymbol{m}, \boldsymbol{h}$ be the vectors of metric and human assessments, respectively. *Concordant* pairs are the pairs of metric assessments that have been ranked in the same order by humans; *discordant* pairs are those ranked in a different order. We define C and D as the number of concordant and discordant pairs, respectively. We also define T_h as the number of pairs only tied in the gold scores, T_m as the number of pairs only tied in the metric scores, and T_{hm} as the number of pairs tied both in gold and metric scores, i.e., the number of correctly predicted ties. The Kendall τ correlation coefficient is defined as follows (Kendall, 1945):

$$\tau = \frac{C - D}{\sqrt{(C + D + T_h)(C + D + T_m)}}. \quad (1)$$

4.2 The acc_{eq}

As noted by Deutsch et al. (2023), Kendall τ penalizes the prediction of ties, but never rewards them, as T_m and T_h are in the denominator, and T_{hm} is not used. This issue was not prominent in the earliest editions of the Metrics Shared Task, where ties in human scores were disregarded, and older metrics rarely produced ties. Currently, instead, it is essential to consider the prediction of ties, especially since human MQM annotations contain a lot of them,[11] and some recently-proposed metrics are designed to output evaluation assessments that resemble MQM (Perrella et al., 2022; Kocmi and Federmann, 2023). For this reason, Deutsch et al. (2023) proposed a measure that mimics the τ coefficient in the way it is computed, but also accounts for correctly predicting ties:

$$\text{acc}_{\text{eq}} = \frac{C + T_{hm}}{C + D + T_h + T_m + T_{hm}}. \quad (2)$$

Differently from Kendall τ, acc_{eq} includes T_{hm} in the numerator, and the denominator encompasses the total number of pairs. Notably, discordant pairs are not subtracted from the numerator, rendering this metric a measure of accuracy, with scores ranging between 0 and 1. In Appendix F, we provide a numerical example of the computation of both Kendall τ and acc_{eq} from the vectors $\boldsymbol{m}$ and $\boldsymbol{h}$.

The acc_{eq} measure, as it stands, would unfairly disadvantage continuous metrics. Indeed, it is extremely infrequent for such metrics to assign the same score to two different translations, meaning that they never predict ties. To address this issue, Deutsch et al. (2023) propose the tie calibration algorithm. In the following section, we briefly illustrate this algorithm and explain why it should not be conducted on the same test set used for the meta-evaluation.

4.3 Tie Calibration

The tie calibration algorithm determines, for each metric, a threshold ϵ such that, given two metric assessments m_1 and m_2, they are tied if $|m_1 - m_2| \leq \epsilon$. Deutsch et al. (2023) propose selecting the ϵ that maximizes acc_{eq} on the same test set used for the metrics meta-evaluation, enabling metrics to output the number of tied scores that best fits the distribution of human ties in the considered test set. This distribution is not stable across test sets (Table 11), and Deutsch et al. (2023) show that ϵ values are not stable either. Nonetheless, they argue that this would not impact the fairness of the evaluation. Unfortunately, our analysis shows that this is not the case. Specifically, despite all metrics' ϵ values being selected on the same test data, we demonstrate that continuous metrics are more flexible to best fit the underlying distribution of human ties, compared to discrete ones, leading to unfairly higher acc_{eq} values.

4.4 Two New Sentinel Metrics

To demonstrate the impact of this phenomenon, we introduce two additional sentinel metrics, i.e., $\text{SENTINEL}_{\text{GEMBA}}$ and $\text{SENTINEL}_{\text{MATESE}}$. GEMBA-MQM (Kocmi and Federmann, 2023) and MaTESe (Perrella et al., 2022) are MT metrics that output discrete scores in the form of MQM quality assessments and participated in WMT23. $\text{SENTINEL}_{\text{GEMBA}}$ and $\text{SENTINEL}_{\text{MATESE}}$ are perturbed versions of GEMBA-MQM and MaTESe, respectively, obtained by adding Gaussian noise – $\mathcal{N}(0, 0.0001)$ – to their predictions. By making their output continuous in the neighborhood of discrete values, we partially fill their gap with continuous metrics, while preventing any two dif-

to discrete metrics, which can take on a limited set of values. Metrics from the COMET family such as COMET, XCOMET-Ensemble, and CometKiwi (Rei et al., 2022b) are continuous, whereas GEMBA-MQM (Kocmi and Federmann, 2023) and MaTESe (Perrella et al., 2022) are examples of discrete metrics.

[10] This is τ_b in Deutsch et al. (2023).

[11] This is also due to the increasing quality of automatic translation, as perfect translations are assigned the same maximum score.

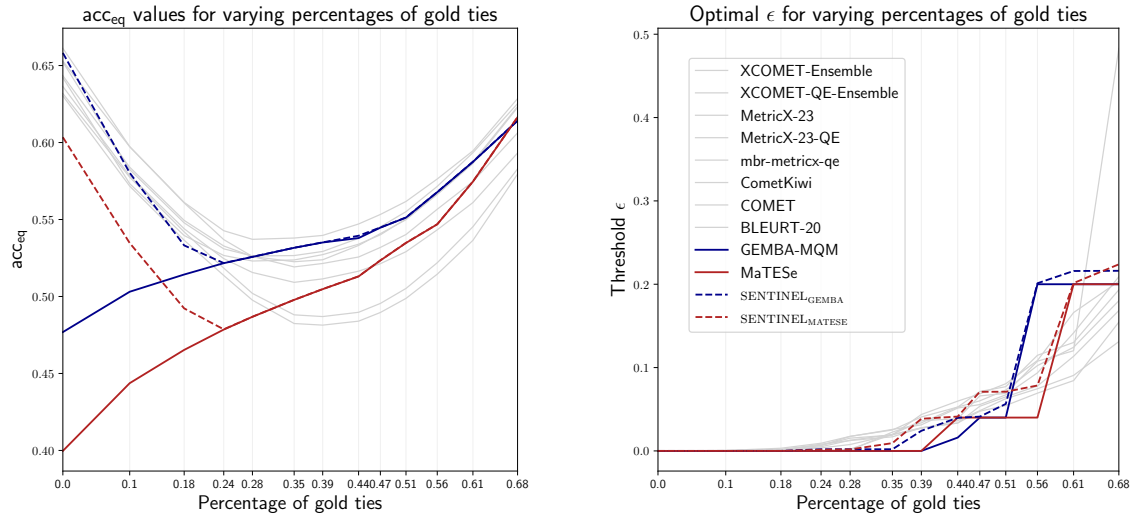


Figure 2: acc_{eq} (left) and optimal ϵ (right) of the considered metrics for varying percentages of human ties in the test dataset, where 0.24 is the percentage of human ties in the entire dataset, obtained when p_t and p_n are both 0. ϵ values have been scaled using min-max scaling. Specifically, for each metric, the minimum ϵ is the optimal ϵ at 0% of human ties, and the maximum is the optimal ϵ at 100%. The language direction is ZH $\rightarrow$ EN. Results concerning all language directions can be found in Appendix G. For each percentage of human ties, we use 5 different seeds to sub-sample the test data. Therefore, the shown acc_{eq} and ϵ, for each metric and percentage of ties, are averaged across 5 different runs.

ferent discrete assessments from inverting their ordering. That is, if two GEMBA-MQM's assessments m_1, m_2 are such that $m_1 > m_2$, this relation is preserved by $\text{SENTINEL}_{\text{GEMBA}}$. In general, we expect a fair meta-evaluation to rank these sentinels on par or below their discrete counterparts. Furthermore, we wish to remark that this solution is sub-optimal compared to metrics that are continuous by design. Indeed, due to the addition of Gaussian noise, the ordering of all $\text{SENTINEL}_{\text{GEMBA}}$ and $\text{SENTINEL}_{\text{MATESE}}$'s assessments in the neighborhood of discrete values is randomized.

To demonstrate that $\text{SENTINEL}_{\text{GEMBA}}$ and $\text{SENTINEL}_{\text{MATESE}}$ can better fit the distribution of human ties compared to their discrete counterparts, we modify such a distribution in the test data. Specifically, we repeatedly sub-sample the test data, such that for each pair of tied human assessments we remove that pair from the test data with a certain probability p_t, and do the same for non-tied pairs, which are removed with probability p_n. We extract 13 samples by assigning various values to p_t and p_n and report the chosen values in Table 12 in Appendix G. As a consequence, each pair (p_t, p_n) represents a different sub-sample of test data, with a different percentage of tied human pairs. Then, for each metric, we select the best ϵ and compute acc_{eq} on each of these samples.

4.5 Results

In Figure 2 (left), we present the acc_{eq} results for a subset of continuous metrics, together with GEMBA-MQM, MaTESe, $\text{SENTINEL}_{\text{GEMBA}}$, and $\text{SENTINEL}_{\text{MATESE}}$. We discuss our results on the WMT23 ZH $\rightarrow$ EN test set, and report results concerning the other language directions, i.e., EN $\rightarrow$ DE and HE $\rightarrow$ EN, in Appendix G. At first glance, it is evident that discrete metrics exhibit a distinct acc_{eq} pattern compared to continuous and sentinel metrics. Notably, at lower percentages of tied human pairs, $\text{SENTINEL}_{\text{GEMBA}}$ and $\text{SENTINEL}_{\text{MATESE}}$ significantly outperform GEMBA-MQM and MaTESe.[12] This discrepancy arises because the tie calibration algorithm selects very small ϵ values, close to 0 for every metric, allowing the number of ties predicted by continuous metrics to potentially drop to 0. Conversely, metrics that yield discrete scores inherently produce a certain number of ties, placing them at a disadvantage, and thus ranking conceptually identical metrics like $\text{SENTINEL}_{\text{GEMBA}}$ and GEMBA-MQM at significantly different positions. Interestingly, in the hypothetical scenario in which there are no tied

[12] It is important to highlight that the range of human tie percentages explored in our analysis is similar to that found in the WMT test sets. Indeed, as shown in Table 11, such percentages range from a minimum of 15.14% to a maximum of 53.35%, observed in the WMT22 EN $\rightarrow$ DE test set.

human pairs in the dataset, SENTINEL$_{\text{GEMBA}}$ would rank second (despite several of its assessments having a random ordering), whereas GEMBA-MQM would be second to last. At increasing percentages of gold ties, instead, the acc$_{\text{eq}}$ values obtained by SENTINEL$_{\text{GEMBA}}$ and SENTINEL$_{\text{MATESE}}$ converge to those of their discrete counterparts. However, this is a limitation of these sentinels' design and does not imply that the evaluation is fair at higher percentages of human ties.

To better investigate the source of unfairness, in Figure 2 (right) we show how the optimal ϵ changes at varying percentages of human ties. As can be seen, continuous metrics' ϵ is dynamically adjusted with heightened sensitivity, contrary to what happens for discrete metrics. Specifically, their ϵ is exactly 0 until the percentage of human ties over all pairs is 39%. Additionally, for MaTESe, it remains constant between 44% and 56%, and between 61% and 68%, and the same happens for GEMBA-MQM between 47% and 51% and between 56% and 68%. In contrast, the values change for all the other metrics in the same intervals, enabling them to better fit the distribution of gold ties found in the test set.

4.5.1 Can we use a held-out set for tie calibration?

We have demonstrated that conducting the tie calibration on the same test set used for the evaluation favors continuous metrics over discrete ones. Nonetheless, this does not necessarily mean using a held-out dataset would ensure a fair meta-evaluation. Indeed, our experiments show that unfairness stems from the different levels of adaptability between continuous and discrete metrics to the distribution of human ties found in the dataset used for tie calibration. Therefore, we expect that using a held-out dataset would still advantage continuous metrics if the distribution of human ties in the held-out resembled that of the test set, and disadvantage them if such a distribution differed from that of the test set. In both cases, continuous metrics' increased adaptability compared to discrete metrics would impair the fairness of the evaluation. To investigate this further, we compute a 80-20 split of the test set to obtain an evaluation set for tie calibration. Then, we repeatedly sub-sample such an evaluation set to modify its distribution of human ties and compute acc$_{\text{eq}}$ on the new test set. The results are shown in Figure 3. We observe that the ranking is unstable at varying percentages of human ties, putting continuous metrics at a disadvantage if the proportion of ties in the evaluation set deviates significantly from that in the test set.

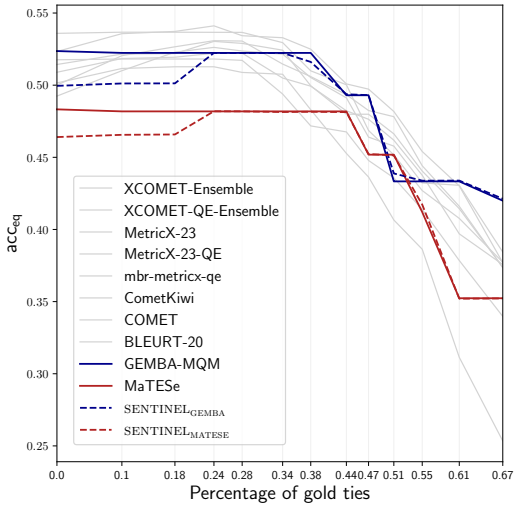


Figure 3: acc$_{\text{eq}}$ of the considered metrics when tie calibration is conducted on a held-out set, derived as a 20% split of the test set, and repeatedly sub-sampled to modify its percentage of tied human scores. The x-axis represents the percentage of ties in the held-out set, while the y-axis represents the acc$_{\text{eq}}$, as computed on the remaining 80% of the test set. The language direction is ZH $\rightarrow$ EN, and results concerning all language directions can be found in Appendix G. The percentage of human ties in the 80% split of the test set is 24%.

5 Conclusion

In this work, we identified two issues with the current meta-evaluation of Machine Translation, as conducted at the Metrics Shared Task of the Conference on Machine Translation. We proposed a suite of sentinel metrics designed to highlight these issues and demonstrate their impact on the metrics rankings, revealing that certain metric categories are unfairly advantaged. Indeed, the *None Grouping* and *System Grouping* strategies favor trained metrics over overlap- and LLM-based ones and the algorithm of tie calibration favors continuous metrics over discrete ones, or vice versa, depending on the percentage of tied assessments in the dataset used for it. Specifically, continuous metrics are favored if the tie calibration is conducted on the same test set used for the evaluation. Finally, we observed a notably high correlation between sentinel metrics and state-of-the-art metrics, raising concerns about their reliability and suggesting that their assessments might be based on spurious correlations present in the training data.

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 has been carried out while Lorenzo Proietti and Alessandro Scirè were enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome.

Limitations

Our analysis recommends grouping translations by their source segment before computing segment-level correlations with human judgments, showing that the rankings derived from the *No Grouping* and *System Grouping* strategies favor certain metric categories and potentially reward metrics for leveraging spurious correlations. However, we recognize that the *Segment Grouping* strategy does not evaluate the ability of metrics to distinguish between higher- and lower-quality translations in absolute terms, that is, independently of their source sentence. We believe this aspect should play a role in the meta-evaluation process, and leave to future work the development of fairer methods to fill this gap. Furthermore, we acknowledge that, due to *Segment Grouping*, each correlation measure is computed on a limited number of data points, i.e., as many as the MT systems that translated each source segment. In this respect, we argue that it would be necessary to investigate the metrics' ranking stability with varying numbers of MT systems, similar to the work of Riley et al. (2024), where they explored MT systems' ranking stability in designing human evaluation studies.

Finally, we acknowledge that we did not provide a clear recommendation regarding a fair option for conducting the tie calibration algorithm. We demonstrated that continuous metrics are favored if selecting the optimal ϵ on the same test set used for the meta-evaluation and that using a held-out dataset would not be fair either. Nonetheless, using a held-out set would at least prevent the distribution of human ties used for tie calibration from being identical to that of the test set, and therefore it should be preferred. In general, we believe that a promising approach might involve studying the meaning of the score deltas of continuous metrics (akin to the work of Kocmi et al. (2024) regarding system-level assessments) and treating as tied all assessments within pre-defined score ranges derived from such deltas. This approach would also enhance the interpretability of MT metrics' assessments.

References

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.

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.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

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. Embed_Llama: Using LLM embeddings for the metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 738–745, Singapore. Association for Computational Linguistics.

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, 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. 2022. 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.

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 Miguel Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *CoRR*, abs/2310.10482.

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.

M. G. Kendall. 1945. The treatment of ties in ranking problems. *Biometrika*, 33:239–51.

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, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies.

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

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2020. On the variance of the adaptive learning rate and beyond. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM) : a framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):455–463.

Matouš Macháček and Ondřej Bojar. 2013. Results of the WMT13 metrics shared task. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 45–51, Sofia, Bulgaria. Association for Computational Linguistics.

Matouš Macháček and Ondřej Bojar. 2014. Results of the WMT14 metrics shared task. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 293–301, Baltimore, Maryland, USA. Association for Computational Linguistics.

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, 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.

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.

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

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

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T.

Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Parker Riley, Daniel Deutsch, George Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus Freitag. 2024. Finding replicable human evaluations via stable ranking probability.

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.

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.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A Official Ranking

In Table 3, we report the official segment-level ranking of WMT23 Metrics Shared Task, including sentinel metrics.

B Training the Sentinel Metrics

The input for the sentinel metrics consists of either the source text ($\text{SENTINEL}_{\text{SRC}}$), candidate translation ($\text{SENTINEL}_{\text{CAND}}$), or reference translation ($\text{SENTINEL}_{\text{REF}}$). Each sentence is tokenized and passed to the XLM-RoBERTa large model, which serves as a feature extractor. Then, we pass the embedding of the [CLS] token to a multi-layer, fully-connected neural network, which outputs the final scalar score. More formally, considering t as the input text for a sentinel metric:

$$\begin{aligned} \boldsymbol{e}_t &= \text{XLM-R}\,(t) \\ \boldsymbol{h}_t^{(1)} &= \text{Dropout}\left(\text{Tanh}\left(W_h^{(1)}\boldsymbol{e}_t + \boldsymbol{b}_h^{(1)}\right)\right) \\ \boldsymbol{h}_t^{(2)} &= \text{Dropout}\left(\text{Tanh}\left(W_h^{(2)}\boldsymbol{h}_t^{(1)} + \boldsymbol{b}_h^{(2)}\right)\right) \\ s_t &= W_o\boldsymbol{h}_t^{(2)} + \boldsymbol{b}_o \end{aligned}$$

Where:

- t is the tokenized input sentence.
- $\boldsymbol{e}_t$ is the [CLS] token embedding at the output of XLM-RoBERTa large.
- $\boldsymbol{h}_t^{(i)}$ represents the output of the i^{th} layer of the fully-connected neural network. Each layer consists of a linear transformation, using weight matrix $W_h^{(i)}$ and bias vector $\boldsymbol{b}_h^{(i)}$, followed by a Tanh activation function and a dropout layer.
- W_o and $\boldsymbol{b}_o$ are the output layer's weight matrix and bias vector, respectively.
- s_t is the output scalar score assigned to sentence t.

Both training phases (i.e., the first, using DA-based human judgments, and the second, using MQM-based ones) employ the same set of hyperparameters, detailed in Table 4.

C Grouping Strategies

In Tables 5, 6, 7, we report the complete set of rankings and Pearson correlations, at the segment level, of the primary submissions to the WMT23 Metrics Shared Task, with sentinel metrics. Sentinel metrics are consistently ranked lower with *Segment Grouping*. Furthermore, in Tables 8, 9, 10, we report the complete set of rankings and Kendall τ correlation coefficients, at the segment level, of the primary submissions to the WMT23 Metrics Shared Task, with sentinel metrics. With Kendall τ as well, sentinel metrics rank lower when *Segment Grouping* is employed. We wish to note that *Segment Grouping* requires the estimation of multiple correlation coefficients, which are then averaged. Consequently, each correlation is measured on a substantially smaller number of data points, compared to *No Grouping* and *System Grouping*. As a result, the number of clusters of statistical significance is reduced. Therefore, one should not

Metric	Avg. corr		EN → DE Pearson		EN → DE acc_{eq}		HE → EN Pearson		HE → EN acc_{eq}		ZH → EN Pearson		ZH → EN acc_{eq}	
XCOMET-Ensemble	1	0.697	1	0.695	1	0.604	1	0.556	1	0.586	1	0.650	1	0.543
MetricX-23	2	0.682	4	0.585	1	0.603	1	0.548	2	0.577	2	0.625	3	0.531
XCOMET-QE-Ensemble*	3	0.681	2	0.679	3	0.588	3	0.498	4	0.554	1	0.647	3	0.533
MetricX-23-QE*	4	0.681	3	0.626	2	0.596	2	0.520	3	0.564	1	0.647	4	0.527
mbr-metricx-qe*	5	0.652	4	0.571	3	0.584	5	0.411	4	0.553	6	0.489	2	0.537
GEMBA-MQM*	6	0.639	6	0.502	5	0.572	5	0.401	3	0.564	7	0.449	5	0.522
MaTESe	7	0.636	5	0.554	9	0.528	4	0.459	5	0.550	5	0.511	12	0.479
CometKiwi*	8	0.632	7	0.475	5	0.569	7	0.387	6	0.544	7	0.442	4	0.525
sescoreX	9	0.628	6	0.519	6	0.563	7	0.385	16	0.484	4	0.536	9	0.499
SENTINEL$_{CAND}$*	10	0.626	5	0.561	6	0.562	10	0.339	16	0.483	3	0.580	14	0.473
cometoid22-wmt22*	11	0.625	8	0.441	4	0.578	9	0.365	12	0.515	6	0.479	7	0.515
KG-BERTScore*	12	0.624	8	0.451	7	0.556	8	0.382	7	0.537	8	0.430	6	0.516
COMET	13	0.622	9	0.432	4	0.574	5	0.401	8	0.532	9	0.396	7	0.514
BLEURT-20	14	0.622	7	0.484	5	0.572	8	0.382	11	0.519	10	0.378	6	0.518
Calibri-COMET22-QE*	15	0.603	9	0.441	12	0.483	6	0.395	13	0.506	7	0.443	10	0.491
Calibri-COMET22	16	0.603	10	0.413	10	0.522	5	0.401	12	0.515	9	0.396	14	0.474
YiSi-1	17	0.600	12	0.366	8	0.542	6	0.395	8	0.529	12	0.290	8	0.504
docWMT22CometDA	18	0.598	11	0.394	7	0.559	10	0.339	14	0.497	11	0.353	10	0.493
docWMT22CometKiwiDA*	19	0.598	8	0.444	8	0.547	12	0.286	15	0.489	9	0.387	10	0.493
prismRef	20	0.593	6	0.516	10	0.518	11	0.319	9	0.528	14	0.183	8	0.504
MS-COMET-QE-22*	21	0.588	13	0.310	8	0.546	12	0.295	14	0.498	10	0.367	9	0.498
BERTscore	22	0.582	13	0.325	9	0.528	10	0.335	12	0.515	13	0.236	9	0.499
mre-score-labse-regular	23	0.558	18	0.111	9	0.530	8	0.378	10	0.522	16	0.145	12	0.481
XLsim	24	0.544	14	0.239	9	0.527	14	0.233	17	0.480	17	0.111	15	0.464
f200spBLEU	25	0.540	14	0.237	9	0.526	14	0.230	19	0.447	18	0.108	13	0.476
MEE4	26	0.539	17	0.202	9	0.529	13	0.256	20	0.441	18	0.105	12	0.480
tokengram_F	27	0.537	16	0.227	10	0.520	14	0.226	18	0.461	20	0.060	11	0.485
chrF	28	0.537	15	0.232	10	0.519	15	0.221	18	0.460	19	0.063	11	0.485
BLEU	29	0.533	17	0.192	10	0.520	15	0.220	20	0.442	17	0.119	14	0.472
prismSrc*	30	0.530	9	0.425	13	0.426	16	0.140	20	0.441	13	0.223	17	0.421
embed_llama	31	0.529	14	0.250	12	0.483	15	0.215	21	0.430	15	0.161	16	0.447
SENTINEL$_{SRC}$*	32	0.512	7	0.469	15	0.231	10	0.334	21	0.428	4	0.540	19	0.240
SENTINEL$_{REF}$	33	0.506	8	0.464	15	0.231	11	0.301	21	0.428	5	0.506	19	0.240
eBLEU	34	0.491	20	−0.011	11	0.512	16	0.131	19	0.445	22	−0.084	14	0.473
Random-sysname*	35	0.463	19	0.064	14	0.409	17	0.041	21	0.428	21	0.018	18	0.381

Table 3: Complete segment-level results for the primary submissions to the WMT 2023 Metrics Shared Task, with sentinel metrics.

Hyperparameter	Value
Optimizer	RAdam (Liu et al., 2020)
Learning Rate	1e-6
Number of Epochs	1
Batch Size	8
Accumulation Steps	2
Dropout	0.1
Dimension of $\boldsymbol{h}_t^{(1)}$	512
Dimension of $\boldsymbol{h}_t^{(2)}$	128

Table 4: Hyperparameters used for both training phases of the sentinel metrics.

focus on the absolute values of the ranks but on their value relative to that of the other metrics. For instance, in Table 9, SENTINEL$_{CAND}$ is ranked 5th out of 19 with *No Grouping*, and 4th out of 11 with *Segment Grouping*. While the absolute value of the rank is lower, in terms of correlation it has moved from the 8th to the 17th position.

D Metrics Pairwise Correlations

In Figures 4, 5, 6, we report the pairwise correlation between a subset of the primary submissions and baselines of WMT23, with the inclusion of sentinel metrics. We use Pearson correlation coefficient with *No Grouping*. State-of-the-art regression-based metrics display a notably high correlation with sentinels. Specifically, the highest correlations are reported by XCOMET-Ensemble, MetricX-23, and their reference-less counterparts. Moderate correlation is also reported between sentinels and baseline metrics such as CometKiwi, COMET, and BLEURT-20. As expected, instead, lexical-based metrics such as BLEU and chrF display close to

Metric	No		Segment		System	
XCOMET-Ensemble	1	0.650	2	0.421	1	0.610
MetricX-23-QE*	1	0.647	4	0.359	1	0.610
XCOMET-QE-Ensemble*	1	0.647	3	0.380	1	0.612
MetricX-23	2	0.625	3	0.373	2	0.580
SENTINEL$_{\text{CAND}}$*	3	0.580	11	0.201	2	0.578
SENTINEL$_{\text{SRC}}$*	4	0.540	14	0.000	3	0.561
sescoreX	4	0.536	7	0.295	5	0.505
MaTESe	5	0.511	6	0.325	6	0.441
SENTINEL$_{\text{REF}}$	5	0.506	14	0.000	4	0.525
mbr-metricx-qe*	6	0.489	1	0.436	7	0.431
cometoid22-wmt22*	6	0.479	4	0.357	6	0.446
GEMBA-MQM*	7	0.449	1	0.434	9	0.378
Calibri-COMET22-QE*	7	0.443	5	0.355	8	0.411
CometKiwi*	7	0.442	3	0.388	9	0.388
KG-BERTScore*	8	0.430	4	0.369	10	0.374
COMET	9	0.396	4	0.364	12	0.345
Calibri-COMET22	9	0.396	7	0.311	11	0.360
docWMT22CometKiwiDA*	10	0.387	6	0.340	13	0.320
BLEURT-20	10	0.378	4	0.371	13	0.330
MS-COMET-QE-22*	11	0.367	7	0.306	14	0.313
docWMT22CometDA	12	0.353	6	0.327	15	0.291
YiSi-1	13	0.290	6	0.329	16	0.237
BERTscore	14	0.236	7	0.309	17	0.186
prismSrc*	15	0.223	13	0.078	16	0.243
prismRef	16	0.183	6	0.332	18	0.135
embed_llama	17	0.161	12	0.138	18	0.139
mre-score-labse-regular	18	0.145	8	0.251	19	0.123
BLEU	19	0.119	11	0.208	20	0.093
XLsim	19	0.111	10	0.218	21	0.069
f200spBLEU	20	0.108	10	0.220	21	0.077
MEE4	20	0.105	9	0.236	21	0.070
chrF	21	0.063	8	0.263	22	0.020
tokengram_F	22	0.060	8	0.262	23	0.015
Random-sysname*	23	0.018	14	0.019	23	0.002
eBLEU	24	−0.084	10	0.219	24	−0.115

Table 5: Segment-level Pearson correlation for the primary submissions to the WMT23 Metrics Shared Task, with sentinel metrics. The language direction is ZH → EN. Starred metrics are reference-free, underlined metrics are baselines, and highlighted metrics are sentinels. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2023), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021).

no correlation with sentinels. Similarly, GEMBA-MQM, a state-of-the-art LLM-based metric that has not been fine-tuned on human assessments, shows lower levels of correlation with the sentinel metrics, compared to the other state-of-the-art metrics.

E Length Bias

In Figures 7,8 we report the relation between metrics assessments and the length of the candidate translation. We concatenate the data from all the three language directions used in the MQM-based evaluation of WMT23, i.e., ZH → EN, EN → DE, and HE → EN. We wish to remind the reader that the meta-evaluation of WMT23 was conducted at the paragraph level for EN → DE, and therefore, the reported candidate lengths are much larger than those in Figure 1, which comprises only ZH → EN. As we can see from the figures, most regression-based metrics, sentinels included, almost never assign very high scores to long translations, even if they are correct. This is in marked contrast to metrics trained with different objectives, such as MaTESe, or not fine-tuned to mimic the human judgment, such as GEMBA-MQM. Indeed, both these metrics assign their highest score to several translations longer than 1200 characters. Notably, there are several metrics whose assessments converge to a very narrow range of values as length increases. For example, BLEURT-20's assessments seem to be confined between approximately 0.4 and 0.8 for translations longer than 1000 characters, and a similar pattern is observed for COMET.

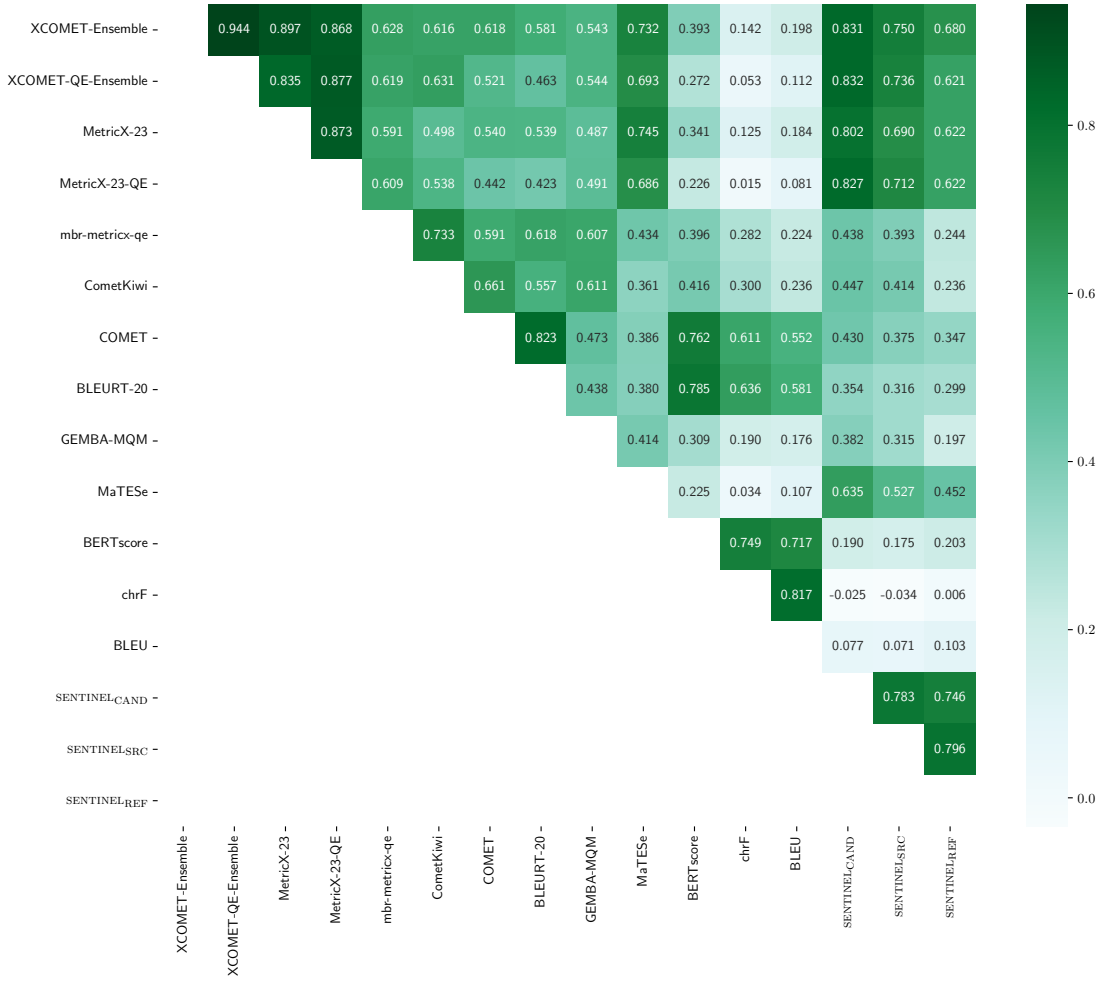


Figure 4: Pairwise correlation between a part of the primary submissions and baselines of WMT23, and sentinel metrics. Correlation is Pearson with *No Grouping*, and the language direction is ZH → EN.

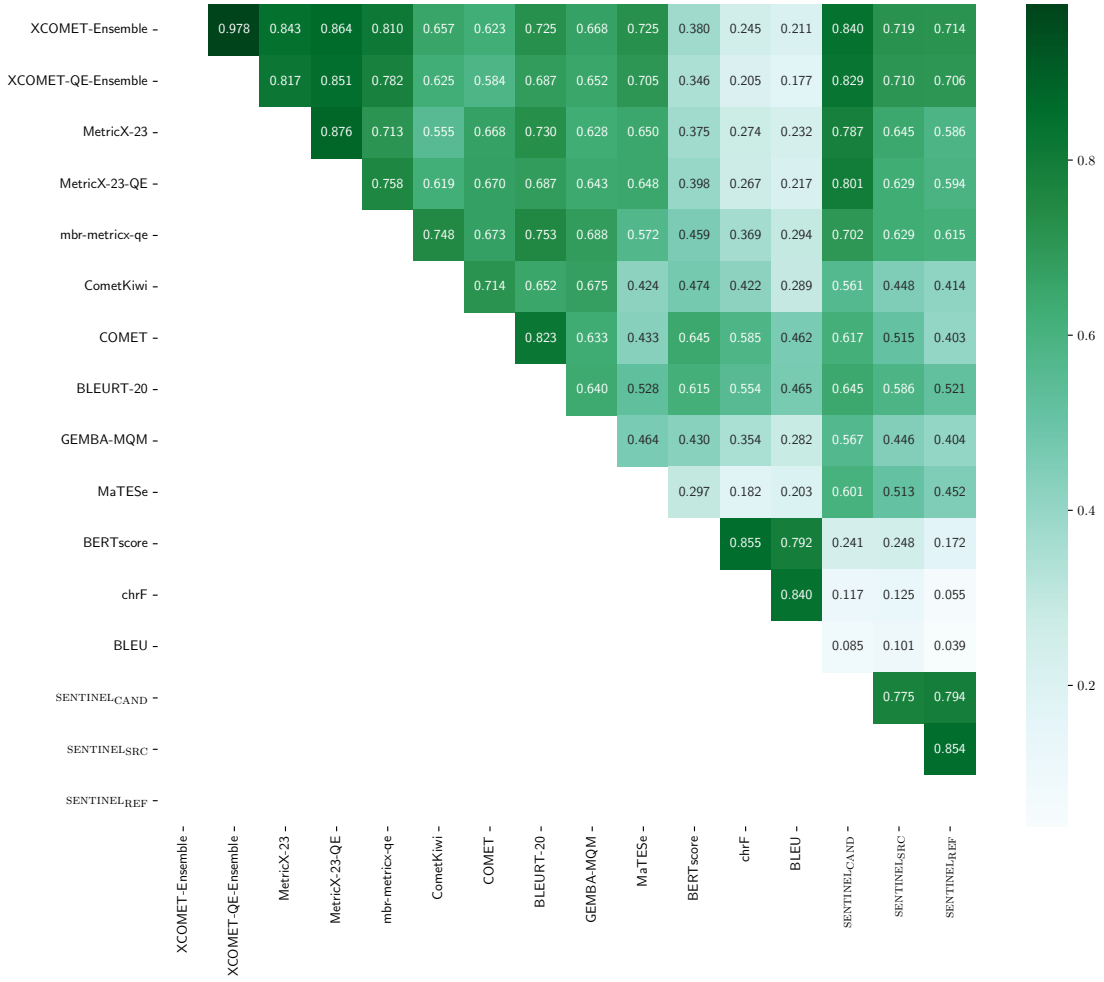


Figure 5: Pairwise correlation between a part of the primary submissions and baselines of WMT23, and sentinel metrics. Correlation is Pearson with *No Grouping*, and the language direction is EN $\rightarrow$ DE.

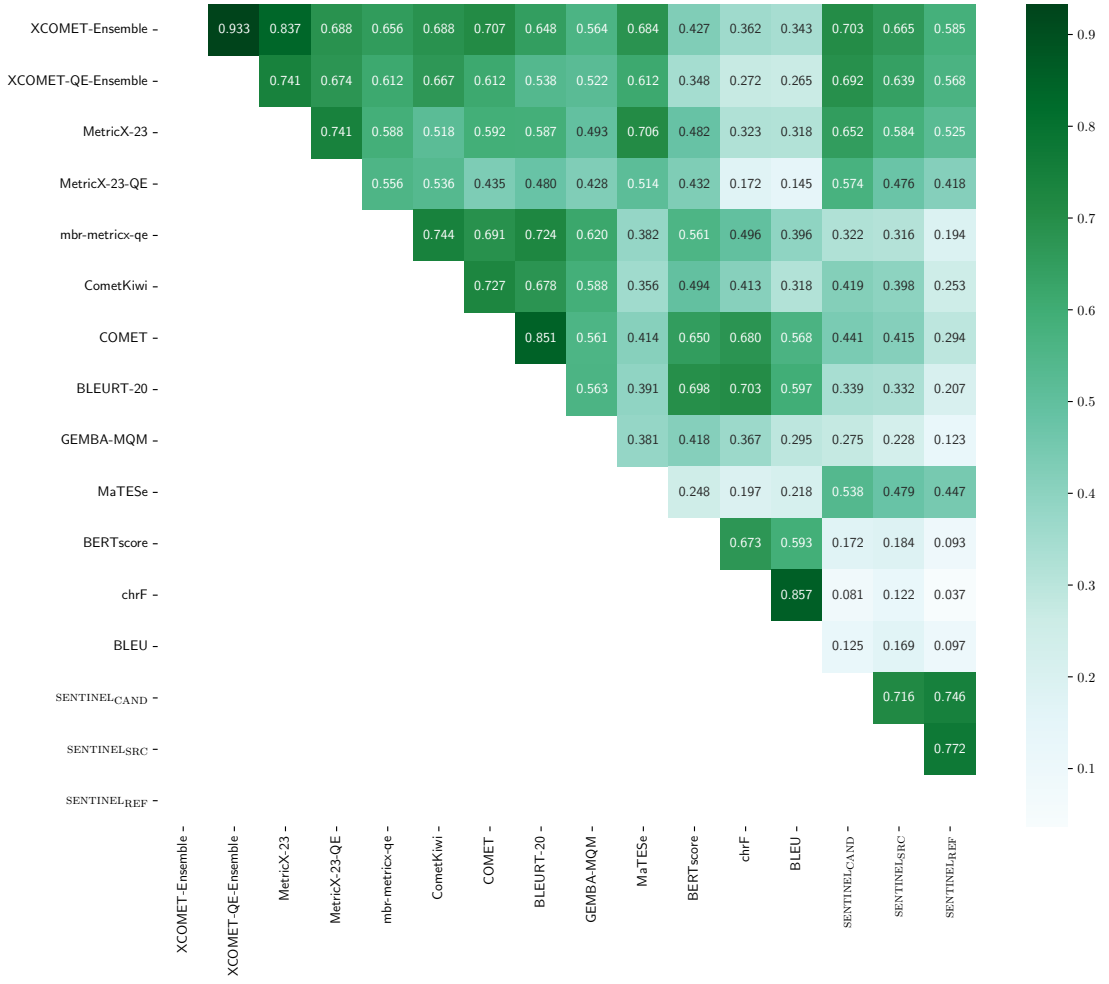


Figure 6: Pairwise correlation between a part of the primary submissions and baselines of WMT23, and sentinel metrics. Correlation is Pearson with *No Grouping*, and the language direction is HE → EN.

Metric	No		Segment		System	
XCOMET-Ensemble	1	0.695	1	0.538	1	0.676
XCOMET-QE-Ensemble*	2	0.679	2	0.507	2	0.658
MetricX-23-QE*	3	0.626	2	0.511	3	0.564
MetricX-23	4	0.585	2	0.507	4	0.547
mbr-metricx-qe*	4	0.571	1	0.543	3	0.551
SENTINEL$_{\text{CAND}}$*	5	0.561	6	0.396	5	0.522
MaTESe	5	0.554	8	0.330	4	0.526
sescoreX	6	0.519	3	0.459	6	0.502
prismRef	6	0.516	7	0.349	4	0.528
GEMBA-MQM*	6	0.502	3	0.482	7	0.446
BLEURT-20	7	0.484	2	0.492	7	0.455
CometKiwi*	7	0.475	3	0.463	7	0.451
SENTINEL$_{\text{SRC}}$*	8	0.469	12	0.000	6	0.502
SENTINEL$_{\text{REF}}$	8	0.464	12	0.000	6	0.492
KG-BERTScore*	8	0.451	4	0.456	8	0.421
docWMT22CometKiwiDA*	9	0.444	5	0.426	9	0.404
cometoid22-wmt22*	9	0.441	2	0.499	9	0.385
Calibri-COMET22-QE*	9	0.441	5	0.432	8	0.414
COMET	9	0.432	2	0.508	10	0.363
prismSrc*	9	0.425	11	0.102	6	0.487
Calibri-COMET22	10	0.413	3	0.477	10	0.370
docWMT22CometDA	11	0.394	3	0.484	11	0.310
YiSi-1	12	0.366	5	0.404	12	0.284
BERTscore	13	0.325	7	0.355	13	0.250
MS-COMET-QE-22*	13	0.310	6	0.400	13	0.241
embed_llama	14	0.250	10	0.242	14	0.180
XLsim	14	0.239	6	0.372	16	0.151
f200spBLEU	14	0.237	7	0.343	14	0.178
chrF	15	0.232	8	0.336	15	0.157
tokengram_F	16	0.227	8	0.340	16	0.153
MEE4	17	0.202	7	0.360	16	0.145
BLEU	17	0.192	9	0.310	17	0.140
mre-score-labse-regular	18	0.111	6	0.376	18	0.087
Random-sysname*	19	0.064	11	0.124	19	−0.015
eBLEU	20	−0.011	8	0.317	19	−0.030

Table 6: Segment-level Pearson correlation for the primary submissions to the WMT23 Metrics Shared Task, with sentinel metrics. The language direction is EN → DE. Starred metrics are reference-free, underlined metrics are baselines, and highlighted metrics are sentinels. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2023), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021).

F Kendall τ and acc$_{\text{eq}}$ Computation Example

In this section, we provide an example of the computation of Kendall τ and acc$_{\text{eq}}$ from two vectors of human and metric scores, i.e., $\boldsymbol{h}$ and $\boldsymbol{m}$ in the following table:

$\boldsymbol{m}$	0.6	0.5	0.4	0.4
$\boldsymbol{h}$	5	3	5	5

For each vector, there are six pairs of assessments. In particular, the pairs of metric assessments are (m_1, m_2), (m_1, m_3), (m_1, m_4), (m_2, m_3), (m_2, m_4), (m_3, m_4).

In Equations 1 and 2, $C = 1$, since the only concordant pair is (m_1, m_2). Indeed, $m_1 > m_2$ and $h_1 > h_2$. $D = 2$, since the pairs $(m_2, m_3), (m_2, m_4)$ are discordant. $T_m = 0$, since there are no pairs tied only in the metric scores. $T_h = 2$, since the pairs $(h_1, h_3), (h_1, h_4)$ are tied only in the human scores. $T_{hm} = 1$, since the remaining pair, i.e., (m_3, m_4), is tied in both human and metric scores. In this example, $\tau = -0.258$ and acc$_{\text{eq}} = 0.333$.

G Ties

In Table 11, we report the percentage of tied human pairs in the datasets used in recent editions of WMT.

In Tables 12, 13, 14, we report the values of p_t and p_n used to sub-sample the ZH → EN, EN → DE, and HE → EN test sets, respectively, to conduct the experiment described in Section 4.4. We also report the corresponding percentage of human ties and total number of pairs, for each sample.

In Figures 9, 10, 11, we report the acc$_{\text{eq}}$ and

Metric	No		Segment		System	
XCOMET-Ensemble	1	0.556	1	0.479	1	0.515
MetricX-23	1	0.548	2	0.441	1	0.509
MetricX-23-QE*	2	0.520	5	0.387	2	0.480
XCOMET-QE-Ensemble*	3	0.498	4	0.397	3	0.458
MaTESe	4	0.459	5	0.373	4	0.408
mbr-metricx-qe*	5	0.411	2	0.448	5	0.362
GEMBA-MQM*	5	0.401	2	0.431	6	0.354
COMET	5	0.401	3	0.421	5	0.367
Calibri-COMET22	5	0.401	4	0.397	5	0.371
YiSi-1	6	0.395	2	0.439	6	0.348
Calibri-COMET22-QE*	6	0.395	6	0.354	5	0.369
CometKiwi*	7	0.387	5	0.375	6	0.353
sescoreX	7	0.385	5	0.370	6	0.352
KG-BERTScore*	8	0.382	5	0.375	7	0.347
BLEURT-20	8	0.382	3	0.418	7	0.344
mre-score-labse-regular	8	0.378	4	0.407	8	0.335
cometoid22-wmt22*	9	0.365	7	0.309	7	0.346
docWMT22CometDA	10	0.339	5	0.379	9	0.294
SENTINEL$_{\text{CAND}}$*	10	0.339	11	0.104	7	0.343
BERTscore	10	0.335	4	0.412	9	0.293
SENTINEL$_{\text{SRC}}$*	10	0.334	13	0.000	7	0.336
prismRef	11	0.319	3	0.428	10	0.276
SENTINEL$_{\text{REF}}$	11	0.301	13	0.000	9	0.299
MS-COMET-QE-22*	12	0.295	9	0.252	10	0.274
docWMT22CometKiwiDA*	12	0.286	7	0.324	11	0.234
MEE4	13	0.256	8	0.291	11	0.222
XLsim	14	0.233	7	0.314	12	0.198
f200spBLEU	14	0.230	8	0.287	12	0.195
tokengram_F	14	0.226	7	0.311	13	0.184
chrF	15	0.221	7	0.308	14	0.179
BLEU	15	0.220	9	0.260	13	0.189
embed_llama	15	0.215	10	0.188	13	0.187
prismSrc*	16	0.140	11	0.100	15	0.150
eBLEU	16	0.131	8	0.280	16	0.104
Random-sysname*	17	0.041	12	0.057	17	0.001

Table 7: Segment-level Pearson correlation for the primary submissions to the WMT23 Metrics Shared Task, with sentinel metrics. The language direction is HE → EN. Starred metrics are reference-free, underlined metrics are baselines, and highlighted metrics are sentinels. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2023), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021).

optimal ϵ for each considered metric, in all three language directions considered at WMT 2023.

In Figure 12, we report the acc_{eq} values of the considered metrics, as computed on a 80% split of the test set. ϵ values have been estimated using a held-out set derived as a 20% split of the entire test set. The held-out set is repeatedly sub-sampled to vary its percentage of tied human scores. Different percentage values are reported on the x-axis.

Metric	No		Segment		System	
XCOMET-Ensemble	1	0.473	2	0.299	1	0.456
XCOMET-QE-Ensemble*	2	0.467	3	0.273	2	0.451
MetricX-23-QE*	3	0.461	4	0.252	2	0.448
GEMBA-MQM*	3	0.457	1	0.365	4	0.416
MetricX-23	4	0.449	3	0.269	3	0.434
mbr-metricx-qe*	5	0.427	2	0.301	5	0.403
cometoid22-wmt22*	5	0.423	4	0.252	4	0.408
SENTINEL$_{\text{CAND}}$*	6	0.404	9	0.148	4	0.410
SENTINEL$_{\text{SRC}}$*	7	0.397	14	0.000	4	0.411
CometKiwi*	7	0.391	3	0.263	6	0.368
Calibri-COMET22-QE*	8	0.386	4	0.241	6	0.366
sescoreX	9	0.375	6	0.217	6	0.367
MaTESe	9	0.371	3	0.271	7	0.345
KG-BERTScore*	10	0.361	4	0.248	8	0.337
SENTINEL$_{\text{REF}}$	11	0.340	14	0.000	7	0.353
COMET	11	0.333	4	0.248	9	0.311
MS-COMET-QE-22*	11	0.332	6	0.213	9	0.311
Calibri-COMET22	12	0.330	6	0.217	9	0.310
BLEURT-20	13	0.310	3	0.261	10	0.288
docWMT22CometKiwiDA*	14	0.299	5	0.234	11	0.265
docWMT22CometDA	15	0.276	5	0.231	12	0.248
prismSrc*	16	0.234	12	0.044	12	0.251
YiSi-1	17	0.220	5	0.231	13	0.196
BERTscore	18	0.180	6	0.216	14	0.156
mre-score-labse-regular	18	0.178	7	0.176	14	0.165
prismRef	19	0.165	5	0.232	15	0.140
embed_llama	20	0.109	11	0.096	16	0.093
XLsim	20	0.101	10	0.140	17	0.080
MEE4	21	0.091	8	0.172	18	0.064
BLEU	21	0.085	9	0.154	18	0.062
f200spBLEU	22	0.068	8	0.165	19	0.042
chrF	23	0.045	7	0.187	20	0.017
tokengram_F	24	0.042	7	0.187	21	0.012
Random-sysname*	25	0.015	13	0.025	22	−0.005
eBLEU	26	−0.041	9	0.156	23	−0.064

Table 8: Segment-level Kendall τ correlation coefficient for the primary submissions to the WMT23 Metrics Shared Task, with sentinel metrics. The language direction is ZH $\rightarrow$ EN. Starred metrics are reference-free, underlined metrics are baselines, and highlighted metrics are sentinels. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2023), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021).

Metric	No		Segment		System	
XCOMET-Ensemble	1	0.546	1	0.380	1	0.530
XCOMET-QE-Ensemble*	2	0.532	2	0.360	2	0.516
MetricX-23-QE*	3	0.509	2	0.357	3	0.487
MetricX-23	3	0.506	2	0.368	3	0.485
sescoreX	4	0.493	3	0.343	4	0.476
mbr-metricx-qe*	4	0.490	1	0.397	4	0.467
GEMBA-MQM*	4	0.482	1	0.399	5	0.449
SENTINEL$_{\text{CAND}}$*	5	0.463	4	0.290	5	0.456
MaTESe	5	0.462	5	0.286	6	0.447
BLEURT-20	6	0.452	2	0.366	7	0.426
SENTINEL$_{\text{SRC}}$*	6	0.443	11	0.000	5	0.462
cometoid22-wmt22*	7	0.422	2	0.362	8	0.398
SENTINEL$_{\text{REF}}$	7	0.418	11	0.000	6	0.437
COMET	7	0.418	2	0.366	9	0.387
Calibri-COMET22	7	0.417	3	0.342	9	0.387
CometKiwi*	8	0.408	3	0.330	9	0.379
Calibri-COMET22-QE*	8	0.406	5	0.279	9	0.379
MS-COMET-QE-22*	9	0.391	5	0.280	10	0.363
KG-BERTScore*	10	0.361	4	0.310	11	0.329
docWMT22CometKiwiDA*	10	0.358	4	0.316	11	0.329
prismRef	11	0.345	6	0.247	11	0.332
docWMT22CometDA	11	0.337	2	0.360	12	0.296
YiSi-1	12	0.280	4	0.297	13	0.250
prismSrc*	12	0.267	10	0.039	12	0.284
BERTscore	13	0.253	5	0.260	14	0.224
MEE4	14	0.225	5	0.271	15	0.190
XLsim	14	0.217	6	0.257	15	0.180
f200spBLEU	15	0.187	6	0.255	16	0.151
chrF	15	0.186	6	0.241	16	0.152
tokengram_F	16	0.183	6	0.245	17	0.149
embed_llama	16	0.182	8	0.163	16	0.150
BLEU	17	0.137	7	0.231	18	0.103
eBLEU	18	0.096	7	0.230	19	0.070
mre-score-labse-regular	18	0.084	5	0.269	19	0.066
Random-sysname*	19	0.033	9	0.081	20	−0.018

Table 9: Segment-level Kendall τ correlation coefficient for the primary submissions to the WMT23 Metrics Shared Task, with sentinel metrics. The language direction is EN $\rightarrow$ DE. Starred metrics are reference-free, underlined metrics are baselines, and highlighted metrics are sentinels. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2023), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021).

Metric	No		Segment		System	
XCOMET-Ensemble	1	0.415	2	0.323	1	0.395
MetricX-23	2	0.401	3	0.302	2	0.382
GEMBA-MQM*	2	0.399	1	0.369	3	0.367
XCOMET-QE-Ensemble*	3	0.374	5	0.276	3	0.358
MetricX-23-QE*	3	0.370	6	0.251	3	0.355
mbr-metricx-qe*	3	0.366	2	0.316	4	0.339
MaTESe	4	0.361	3	0.302	4	0.341
COMET	5	0.350	3	0.309	5	0.327
Calibri-COMET22	6	0.348	4	0.284	6	0.324
BLEURT-20	6	0.344	4	0.295	6	0.320
sescoreX	6	0.342	4	0.285	6	0.320
CometKiwi*	7	0.338	6	0.238	6	0.323
Calibri-COMET22-QE*	7	0.336	7	0.230	6	0.322
YiSi-1	7	0.333	2	0.325	7	0.303
mre-score-labse-regular	7	0.328	4	0.284	7	0.300
KG-BERTScore*	8	0.322	6	0.242	7	0.304
cometoid22-wmt22*	9	0.310	7	0.216	7	0.301
prismRef	9	0.302	3	0.309	8	0.273
BERTscore	10	0.295	4	0.298	9	0.266
docWMT22CometDA	11	0.278	5	0.270	10	0.249
MS-COMET-QE-22*	12	0.261	9	0.174	10	0.249
SENTINEL$_{SRC}$*	13	0.243	12	0.000	10	0.247
SENTINEL$_{CAND}$*	13	0.243	11	0.049	10	0.249
XLsim	13	0.233	7	0.228	11	0.211
MEE4	13	0.231	7	0.221	11	0.202
docWMT22CometKiwiDA*	14	0.227	7	0.229	12	0.192
SENTINEL$_{REF}$	15	0.210	12	0.000	11	0.214
tokengram_F	15	0.207	7	0.228	13	0.175
chrF	16	0.204	7	0.224	14	0.171
f200spBLEU	17	0.193	7	0.219	15	0.162
BLEU	18	0.184	8	0.205	16	0.157
embed_llama	18	0.174	10	0.147	16	0.151
eBLEU	19	0.166	8	0.209	17	0.141
prismSrc*	19	0.164	11	0.043	14	0.169
Random-sysname*	20	0.027	11	0.033	18	0.002

Table 10: Segment-level Kendall τ correlation coefficient for the primary submissions to the WMT23 Metrics Shared Task, with sentinel metrics. The language direction is HE $\rightarrow$ EN. Starred metrics are reference-free, underlined metrics are baselines, and highlighted metrics are sentinels. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2023), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021).

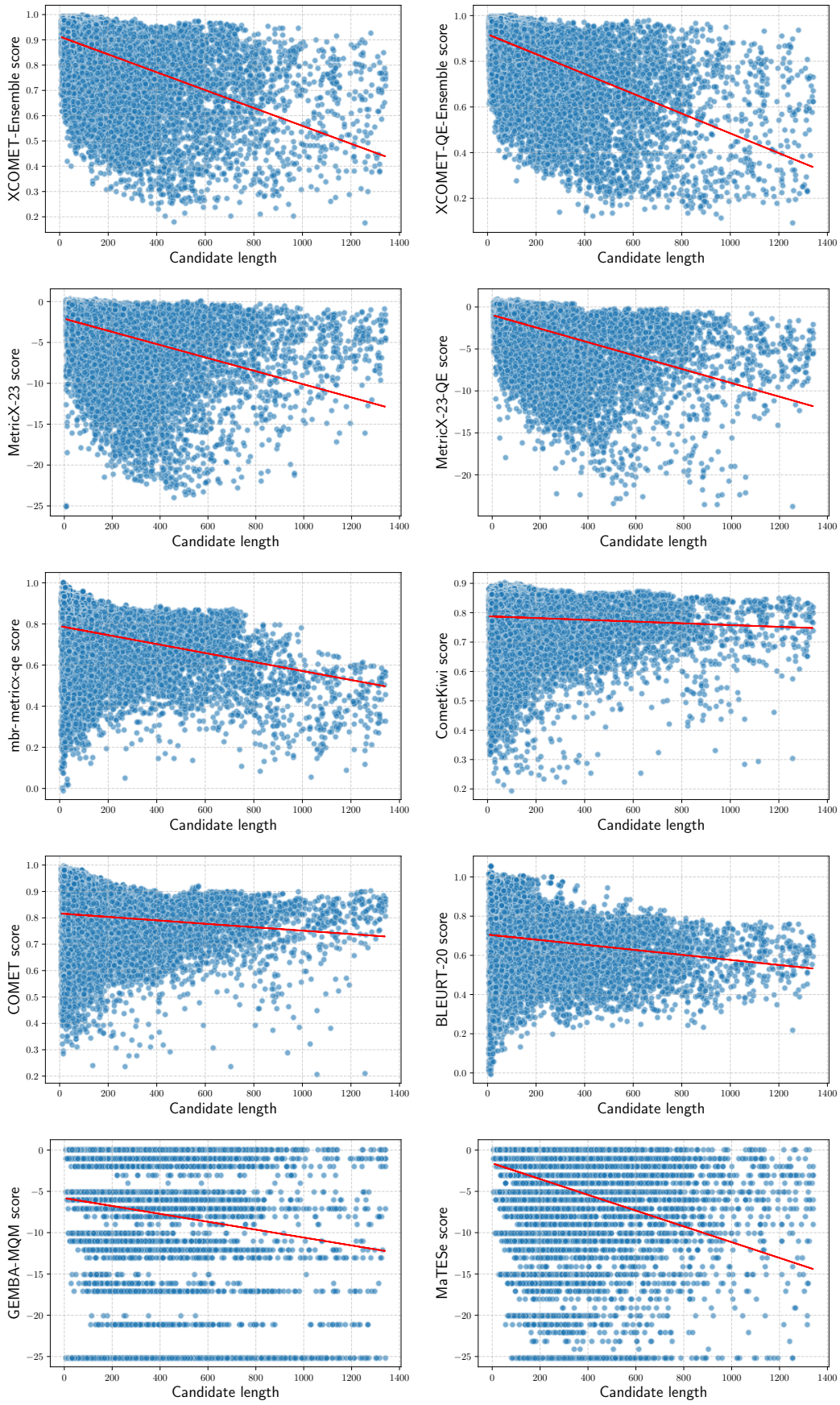


Figure 7: Metric assessments over translation length for a subset of the metrics that participated in WMT23. The red line represents the least-squares regression.

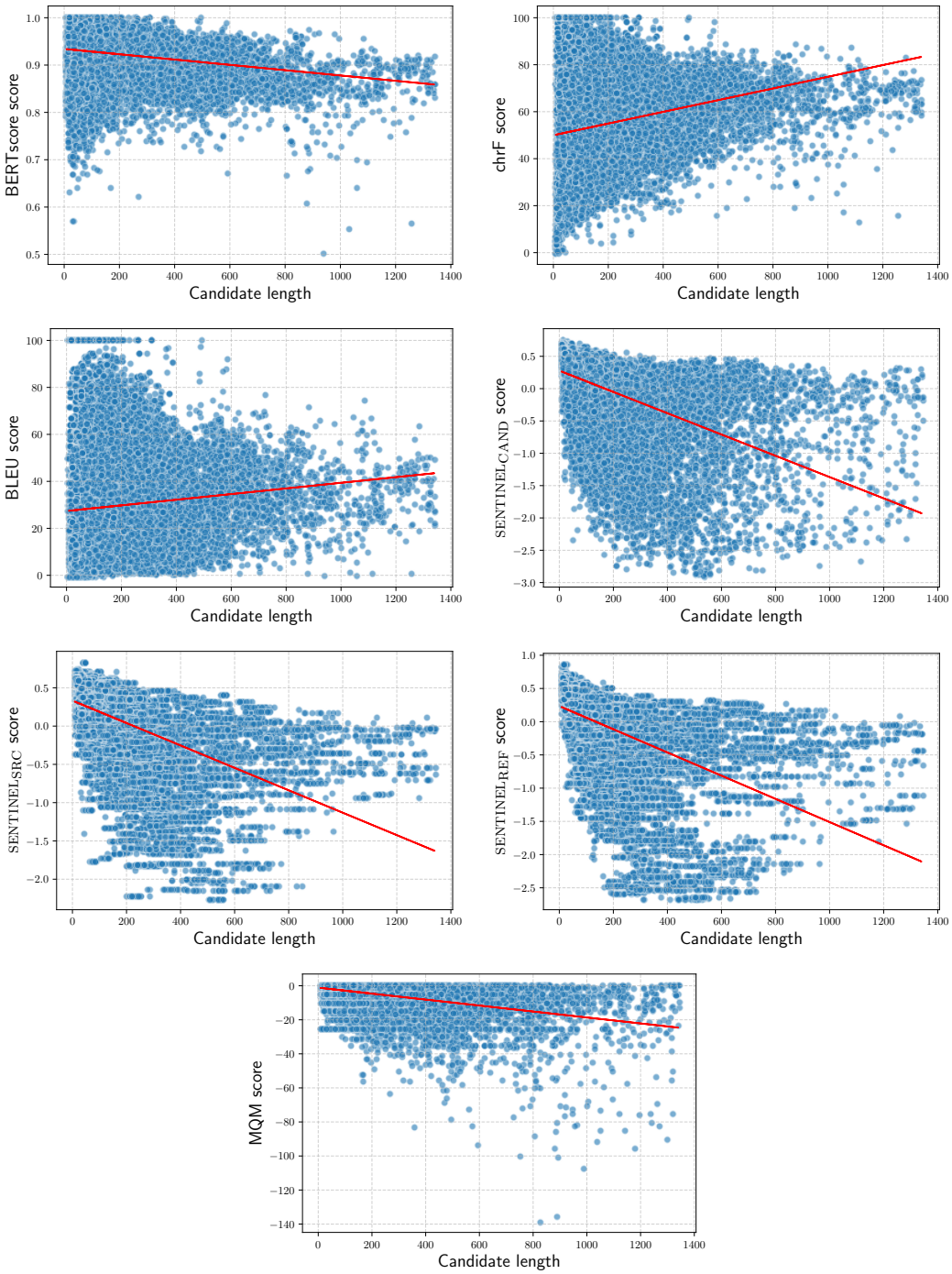
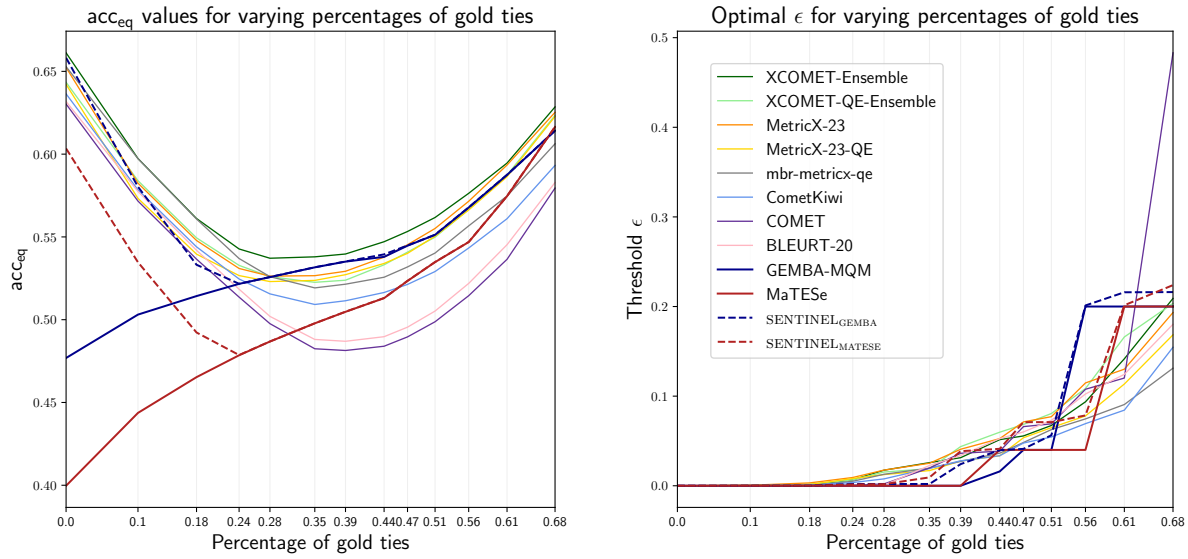
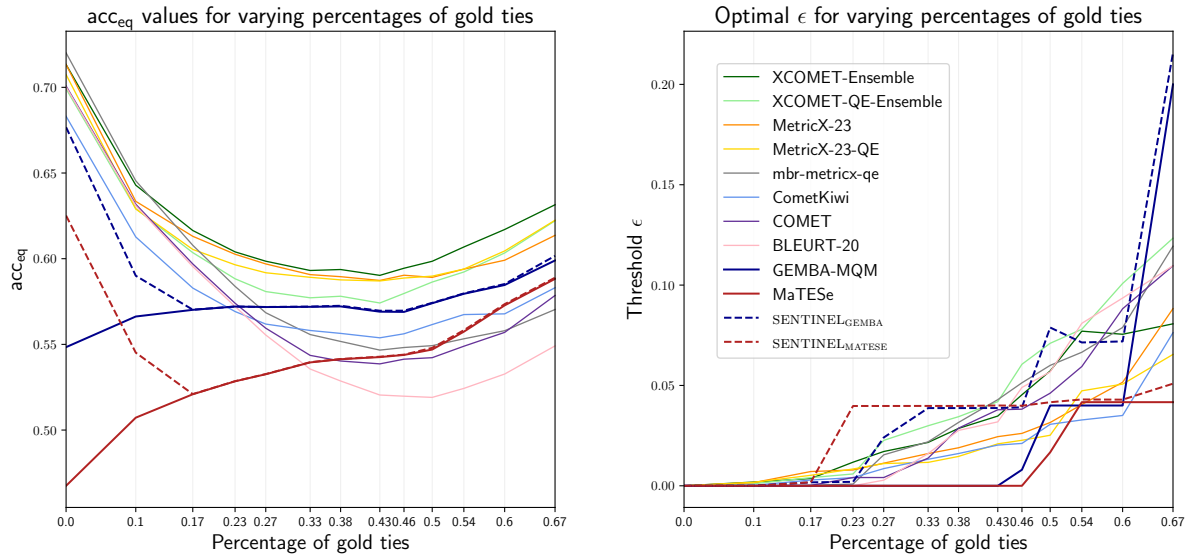


Figure 8: Metric assessments over translation length for a subset of the metrics that participated in WMT23, together with sentinel metrics. The red line represents the least-squares regression.



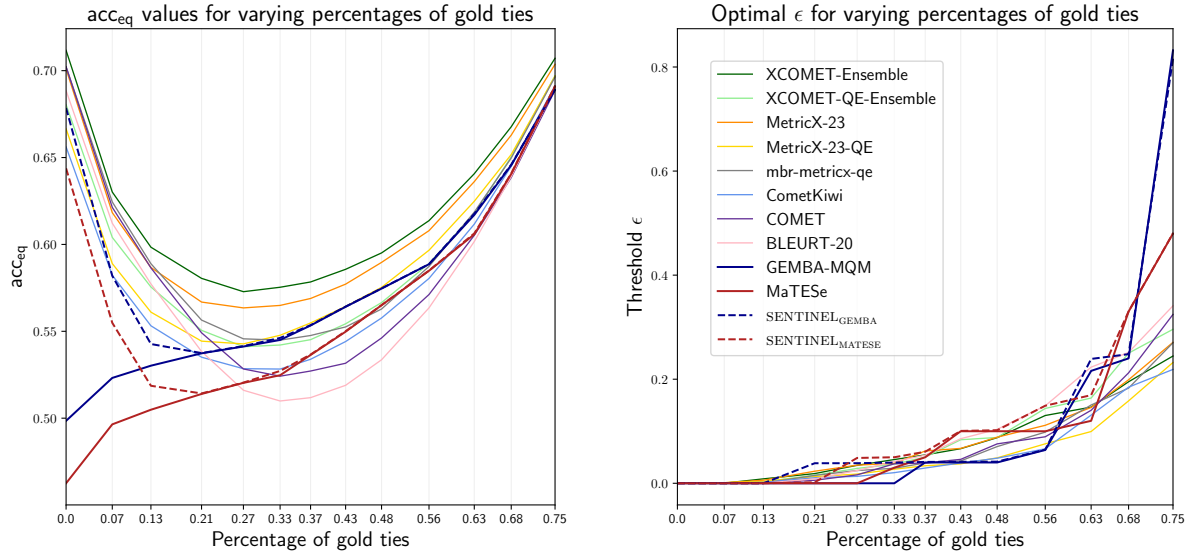
(a) (b)

Figure 9: acc_{eq} (a) and optimal ϵ (b) of the considered metrics for varying percentages of human ties in the test dataset (0.24 is the percentage of human ties in the entire dataset, obtained when p_t and p_n are both 0). ϵ values have been scaled using min-max scaling. Specifically, for each metric, the minimum ϵ is the optimal ϵ at 0% of human ties, and the maximum is the optimal ϵ at 100%. The language direction is ZH $\rightarrow$ EN. For each percentage of human ties, we use 5 different seeds to sub-sample the test data. Therefore, the shown acc_{eq} and ϵ, for each metric and percentage of ties, are averaged across 5 different runs.



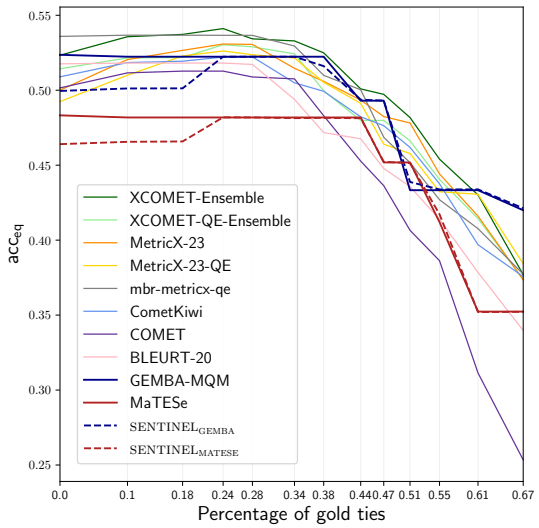
(a) (b)

Figure 10: acc_{eq} (a) and optimal ϵ (b) of the considered metrics for varying percentages of human ties in the test dataset (0.23 is the percentage of human ties in the entire dataset, obtained when p_t and p_n are both 0). ϵ values have been scaled using min-max scaling. Specifically, for each metric, the minimum ϵ is the optimal ϵ at 0% of human ties, and the maximum is the optimal ϵ at 100%. The language direction is EN $\rightarrow$ DE. For each percentage of human ties, we use 5 different seeds to sub-sample the test data. Therefore, the shown acc_{eq} and ϵ, for each metric and percentage of ties, are averaged across 5 different runs.

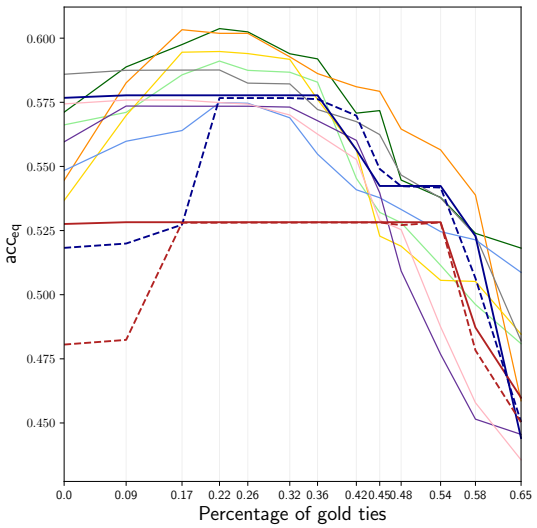


(a) (b)

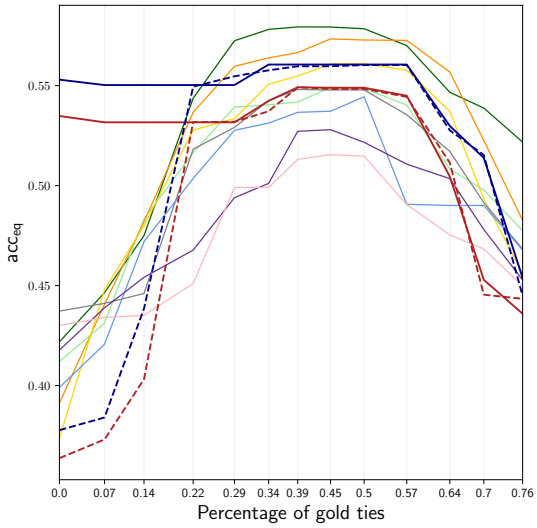
Figure 11: acc_{eq} (a) and optimal ϵ (b) of the considered metrics for varying percentages of human ties in the test dataset (0.43 is the percentage of human ties in the entire dataset, obtained when p_t and p_n are both 0). ϵ values have been scaled using min-max scaling. Specifically, for each metric, the minimum ϵ is the optimal ϵ at 0% of human ties, and the maximum is the optimal ϵ at 100%. The language direction is HE $\rightarrow$ EN. For each percentage of human ties, we use 5 different seeds to sub-sample the test data. Therefore, the shown acc_{eq} and ϵ, for each metric and percentage of ties, are averaged across 5 different runs.



(a) The language pair is ZH → EN. The percentage of human ties in the 80% split of the test set is 24%.



(b) The language pair is EN → DE. The percentage of human ties in the 80% split of the test set is 23%.



(c) The language pair is HE → EN. The percentage of human ties in the 80% split of the test set is 42%.

Figure 12: acc_{eq} of the considered metrics when tie calibration is conducted on a held-out set, derived as a 20% split of the test set, and repeatedly sub-sampled to modify its percentage of tied scores. The x-axis represents the percentage of ties in the held-out set, while the y-axis represents the acc_{eq}, as computed on the remaining 80% of the test set. For each percentage of human ties, we use 5 different seeds to sub-sample the held-out set. Therefore, the shown acc_{eq} for each metric and percentage of ties is averaged over 5 different runs.

	2020	2021	2022	2023
EN → DE	15.14	44.62	53.35	23.11
ZH → EN	17.01	30.31	41.55	24.03
EN → RU	–	53.24	44.42	–
HE → EN	–	–	–	42.84

Table 11: Percentage of tied pairs in the MQM data released over different years at the Metrics Shared Task (or by Freitag et al. (2021a), for 2020), and regarding different translation directions.

p_t	1.00	0.65	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
p_n	0.00	0.00	0.00	0.00	0.20	0.40	0.50	0.60	0.65	0.70	0.75	0.80	0.85
%	0	10	18	24	28	35	39	44	47	51	56	61	68
#	93890	104304	114664	123585	104888	85969	76522	67237	62624	57948	53110	48491	43730

Table 12: p_t is the probability of removing a tied human pair, and p_n is that of removing a non-tied human pair. The considered test set is WMT23 ZH $\rightarrow$ EN. Each column, i.e., each pair (p_t, p_n), represents a sub-sample of the test set, in which tied and non-tied pairs have been removed with such probabilities. The third row contains the percentage of tied human pairs over all pairs, as a result of the sub-sampling. The last row contains the total number of pairs remaining in the test set after the sub-sampling.

p_t	1.00	0.65	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
p_n	0.00	0.00	0.00	0.00	0.20	0.40	0.50	0.60	0.65	0.7	0.75	0.80	0.85
%	0	10	17	23	27	33	38	43	46	50	54	60	67
#	23343	25803	28236	30360	25694	21021	18689	16353	15184	14014	12899	11698	10493

Table 13: p_t is the probability of removing a tied human pair, and p_n is that of removing a non-tied human pair. The considered test set is WMT23 EN $\rightarrow$ DE. Each column, i.e., each pair (p_t, p_n), represents a sub-sample of the test set, in which tied and non-tied pairs have been removed with such probabilities. The third row contains the percentage of tied human pairs over all pairs, as a result of the sub-sampling. The last row contains the total number of pairs remaining in the test set after the sub-sampling.

p_t	1.0	0.90	0.80	0.65	0.50	0.35	0.20	0.00	0.00	0.00	0.00	0.00	0.00
p_n	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.40	0.55	0.65	0.75
%	0	7	13	21	27	33	38	43	48	56	62	68	75
#	36561	39254	42038	46202	50272	54435	58516	63960	56679	49315	43918	40145	36530

Table 14: p_t is the probability of removing a tied human pair, and p_n is that of removing a non-tied human pair. The considered test set is WMT23 HE $\rightarrow$ EN. Each column, i.e., each pair (p_t, p_n), represents a sub-sample of the test set, in which tied and non-tied pairs have been removed with such probabilities. The third row contains the percentage of tied human pairs over all pairs, as a result of the sub-sampling. The last row contains the total number of pairs remaining in the test set after the sub-sampling.