Results of the WMT21 Metrics Shared Task: Evaluating Metrics with Expert-based Human Evaluations on TED and News Domain

Markus Freitag(1), **Ricardo Rei**(2,3,4), **Nitika Mathur**(5,6), **Chi-kiu Lo**(7),
Craig Stewart(2), **George Foster**(1), **Alon Lavie**(2), **Ondřej Bojar**(8)
(1)Google Research (2)Unbabel (3)INESC-ID (4)Instituto Superior Técnico
(5)The University of Melbourne (6)Oracle Digital Assistant
(7)National Research Council Canada (NRC-CNRC) (8)Charles University

Abstract

This paper presents the results of the WMT21 Metrics Shared Task. Participants were asked to score the outputs of the translation systems competing in the WMT21 News Translation Task with automatic metrics on two different domains: news and TED talks. All metrics were evaluated on how well they correlate at the system- and segment-level with human ratings. Contrary to previous years' editions, this year we acquired our own human ratings based on expert-based human evaluation via Multidimensional Quality Metrics (MQM). This setup had several advantages: (i) expert-based evaluation has been shown to be more reliable, (ii) we were able to evaluate all metrics on two different domains using translations of the same MT systems, (iii) we added 5 additional translations coming from the same system during system development. In addition, we designed three challenge sets that evaluate the robustness of all automatic metrics. We present an extensive analysis on how well metrics perform on three language pairs: English→German, English→Russian and Chinese→English. We further show the impact of different reference translations on reference-based metrics and compare our expert-based MQM annotation with the DA scores acquired by WMT.

1 Introduction

The metrics shared task[1] has been a key component of WMT since 2008, serving as a way to validate the use of automatic MT evaluation metrics and driving the development of new metrics. We evaluate reference-based automatic metrics that score MT output by comparing the MT with a reference translation generated by human translators, who are instructed to translate "from scratch" without post-editing from MT. In addition, we also invited submissions of reference-free metrics (quality estimation metrics or QE metrics) that compare MT outputs directly with the source segments. All metrics are evaluated based on their agreement with human rating when scoring MT systems and human translations at the system or sentence level. This year, we implemented several changes to the methodology that was followed in previous years editions of the task:

- **Expert-based human evaluation** This year, we collected our own human ratings for selected language pairs (en→de, en→ru, zh→en) from professional translators via MQM (Lommel et al., 2014). As shown before (Freitag et al., 2021), this produces more reliable[2] scores when compared to the DA-based human ratings acquired by the WMT News-Translation task. This step was necessary as Freitag et al. (2021) suggested that some automatic metrics already outperform (taking MQM as the golden standard) the DA-based human ratings that were usually used in the past for the metrics task and thus the DA-based ground-truth may be of lower quality than some of our submissions.

- **Additional Training Data** We encouraged the participants to further fine tune or test their metrics on the already existing MQM annotations for newstest2020 (Freitag et al., 2021)[3].

- **Additional domain** Since we collected our own human ratings, we were also able to expand the domain of the test sets beyond news and evaluate the performance of the metrics on translations of the same MT systems on TED talks, in order to test the generalization power of metrics.

[1] http://www.statmt.org/wmt21/metrics-task.html

[2] DA is unreliable for high-quality MT output; ranks human translations lower than MT; correlates poorly with metrics. Expert-based MQM ranks human translations higher than MT and correlates generally much better with automatic metrics.

[3] https://github.com/google/wmt-mqm-human-evaluation

Proceedings of the Sixth Conference on Machine Translation (WMT), pages 733–774
November 10–11, 2021. ©2021 Association for Computational Linguistics

- **Additional MT systems** One use case for automatic metrics is choosing the better among different model versions of the same MT system during system development. To address this scenario, in addition to the WMT submissions and online systems, we added extra development systems to the set of MT systems on which we evaluated the metrics.

- **Additional challenge sets** We generated three challenge sets containing specific translation errors that are believed to be challenging for automatic MT evaluation metrics to identify. These challenge sets test metrics robustness on several different phenomena such as sentiment polarity, antonym replacement, named entities, among others.

- **Designated primary metrics** Participants had to designate a single metric as their primary submission for each track (reference-free and unconstrained). Other submissions were permitted, but only the primary metric is included in the official main results.

- **Accuracy for ranking system pairs** We calculate a joint score across all language pairs and adopt the pairwise accuracy score for ranking system pairs to generate the final metric ranking (Kocmi et al., 2021).

Our main findings are:

- WMT direct assessment (DA) scores generally correlate poorly with MQM scores, and exhibit weaker preference for human translations compared to machine output. In particular for English→German and Chinese→English, the two human evaluations methodologies produce very different rankings (Tables 16 and 18). In both language pairs, DA ranks the human translations below many MT systems, demonstrating again that expert-based evaluation is needed to generate a reliable ground truth for metric development for high quality language pairs.

- The majority of automatic metrics correlate better with MQM than the DA scores from WMT. This confirms the findings of Freitag et al. (2021) that automatic metrics are already more reliable than non-expert human evaluations. A metrics task with ground truth ratings acquired by non-experts would consequently not be very helpful.

- The performance of many metrics largely varies depending on the underlying domain (being either news or TED talks), resulting in distinct clusters of winning metrics for these two domains. All metrics of the winning cluster on the news domain show lower correlation with human ratings when switching to the TED talks domain (Table 8). Lower ranked metrics are more robust and can sometimes even improve the correlation to humans on the TED domain.

- Trainable embedding-based metrics are typically better at rating and correctly ranking (with respect to MQM golden truth) human-generated translations. (Table 8).

- Reference-free metrics, in particular COMET-QE and OpenKiwi perform very well when human translations are included in the setup. Nevertheless, once we focus on MT output only, reference-free metrics perform worse compared to reference-based metrics (Table 8).

- Reference-based metrics performance is significantly worse when reference translations contain major errors (Table 13).

- Some metrics are more robust than others when presented with alternate reference translations (Table 14). It is unclear so far what characterizes a good reference translation in addition to the clear requirement of fidelity of the translation to the source.

- When counting top performances across different language pairs, granularities, and test conditions (Table 12), three embedding-based metrics—C-SPECPN, BLEURT-20[4], and COMET-MQM_2021— emerge as distinctly better than the others, especially at the segment level and when rating human translations. Reference-free metrics are also relatively good at rating human translations, but under-perform at segment-level. Metric performance is distributed more evenly on

[4] BLEURT-20 denotes the new retrained version of BLEURT which is different from last years BLEURT submission (Sellam et al., 2020)

system-level tasks, especially when the test set is out-of-domain.

- Most metrics struggle to accurately penalize translations with errors in reversing negation or sentiment polarity (Table 9).

- Of the 14 linguistically motivated categories represented in the challenge sets, high-performing metrics have lower correlations for *Subordination* and *Named Entities and Terminology* (Tables 10 and 11).

- MQM annotations on TED data, both between annotation setups (Google and Unbabel) and between annotators themselves, show relatively low levels of agreement. However, we note that many of the system rankings remain relatively consistent; critically we note that the human reference comes out on top in both setups and that resulting metrics ranking is not significantly affected. This indicates that whilst MQM is an attractive framework for evaluation, the annotation task itself is still subject to human disagreement, especially on challenging content. TED talks in particular are highly specialized and ambiguous, presenting a unique challenge for annotators and evaluation.

2 Data

Similar to the previous years' editions, the source, reference texts, and MT system outputs for the metrics task are mainly derived from the WMT21 News Translation Task. This year, we expand the domain and evaluate the same MT systems on an additional out-of-domain data – TED talks, for our three primary language pairs: English→German, English→Russian and Chinese→English. In addition to the MT system outputs from the WMT evaluation campaign, we added translations from five additional MT systems that represent different versions of the same system during system development.

2.1 WMT Test Sets

The Newstest2021 set contains between 1000 and 2000 segments for each translation direction. All test sets are from the news domain. The reference translations provided for Newstest2021 were created in the same translation direction as the MT systems. We have two reference translations for English→Russian and Chinese→English and four reference translations for English→German. For more details regarding the news test sets, we refer the reader to the WMT21 news translation task findings paper.

2.2 TED Talks Test Suite

A long standing question about automated MT evaluation metrics has been whether metrics *generalize and perform well across domains*. In the past, metrics were mostly tested on news translation evaluation. The WMT2016 metrics shared task (Bojar et al., 2016) experimented on the IT and medical domains but the number of MT systems involved were small (2-10 in each translation direction). Thus, there was insufficient statistical evidence collected for a detailed analysis on how well metrics perform in different domains.

In an attempt to conduct a detailed analysis on the robustness of metrics when evaluating translations in a domain other than news, we generated and provided an additional test suite for translation by the MT systems participating in the news translation task, consisting of transcriptions of TED talks. The TED domain is quite different from the news domain, particularly in its more informal and disfluent language style, yet it covers a wide variety of topics and vocabularies.

The TED talk transcripts translation test set was extracted from OPUS[5] based on the corpus released by Reimers and Gurevych (2020). The English TED talk transcripts were translated by volunteers into multiple languages. To minimize the problem of translationese as the source for the Chinese→English part of the test suite, we had a first-language Chinese speaker select talks with Chinese translations that were judged to be natural-sounding in Chinese. (Unfortunately, there are still some problems in the translation quality for the Chinese→English part of the test suite which we will further discuss in Section 8.1.1.) Then, the same talks were extracted from the corpus to create the English→German and English→Russian parts of the test suite, where the translation was already available in the corpus and the quality of the translation was approved by professional translators. Table 1 shows the basic statistics of the TED talks test suite.

[5] https://opus.nlpl.eu/TED2020.php

language	#talks	#source sent.
en→de	6	606
en→ru	8	684
zh→en	9	843

Table 1: Statistics of the TED talks test suite.

2.3 Additional MT Output

One major use case for automatic metrics is choosing among different versions of the same system during system development. We translated all test sets for all language pairs with five different versions of the same system which we call *metricsystem{1,..,5}*. The underlying NMT models are trained on unconstrained training data and the model variations include baseline models, fine-tuned models and models considering document context. As we will see, the quality performance of these systems and their relative rankings can be quite different depending on the language pair, as these were not trained to yield the highest performance on the news or TED domain.

3 MQM Human Evaluation

Automatic metrics are usually evaluated by measuring correlations with human ratings. The quality of the underlying human ratings is critical and recent findings (Freitag et al., 2021) have shown that crowd-sourced human ratings are not reliable for high quality MT output. Furthermore, an evaluation schema based on MQM (Lommel et al., 2014) which requires explicit error annotation is preferable to an evaluation schema that only asks raters for a single scalar value per translation. Contrarily to the previous versions of the WMT metrics task, for our primary evaluation this year, we decided not to use the crowd-sourced DA human ratings from the WMT News Translation task, and conducted our own MQM-based human evaluation on a subset of submissions and a subset of language pairs that are most interesting for evaluating current metrics. This not only had the advantage of more reliable ratings for a subset of language pairs, but also gave us the opportunity to run the same human evaluation on a different domain (TED talks) on output generated by the same MT systems, in order to test the generalization capabilities of the metrics.

MQM is a general framework that provides a hierarchy of translation errors which can be tailored to specific applications. Google and Unbabel sponsored the human evaluation for this year's metrics task for a subset of language pairs using either professional translators (English→German, Chinese→English) or trusted and trained raters (English→Russian). The error annotation typology and guidelines used by Google's and Unbabel's annotators differs slightly and is described in the following two sections.

3.1 English→German and Chinese→English

Annotations for English→German and Chinese→English were sponsored and executed by Google, using 23 professional translators (14 for English→German, 9 for Chinese→English) with access to the full document context. Instead of assign a scalar value to each translation, annotators were instructed to "just" label error spans within each segment in a document, paying particular attention to document context. Each error was highlighted in the text, and labeled with an error category and a severity. To temper the effect of long segments, we imposed a maximum of five errors per segment, instructing raters to choose the five most severe errors for segments containing more errors. Segments that are too badly garbled to permit reliable identification of individual errors are assigned a special *Non-translation* error. Error severities are assigned independent of category, and consist of *Major*, *Minor*, and *Neutral* levels, corresponding respectively to actual translation or grammatical errors, smaller imperfections, and purely subjective opinions about the translation. Since we are ultimately interested in scoring segments, we adopt the weighting scheme shown in Table 2, in which segment-level scores can range from 0 (perfect) to 25 (worst). The final segment-level score is an average over scores from all annotators. For more details, exact annotator instructions and a list of error categories, we refer the reader to Freitag et al. (2021) as the exact same setup was used for the WMT21 metrics task.

Severity	Category	Weight
Major	Non-translation	25
	all others	5
Minor	Fluency/Punctuation	0.1
	all others	1
Neutral	all	0

Table 2: Google's MQM error weighting.

3.2 English→Russian

Annotation for English→Russian was performed by Unbabel who used a single professional native language annotator with several years of translation error experience based on variations of the MQM framework (Lommel et al., 2014). For this task, Unbabel provided a proprietary variant of MQM, specifically tailored for Russian language annotation. In a manner similar to the Google annotation, the annotator was given full document context and instructed to highlight spans of errors according to the categories specified in the typology. As with the Google annotation, the annotator was also instructed to indicate error severity. The Unbabel severity options differ slightly from that of Google in that we also specify a 'critical' error severity and do not specify a 'neutral' category. Additionally, in the Unbabel typology, all error categories are weighted equally within each severity level.

MQM scores at a segment level are calculated by summing the number of errors in the segment in each severity and applying a severity weight as described in Table 3. In contrast to the Google scheme, Unbabel does not impose a limit on the number of errors in a segment. We do, however, apply a normalization of the score by segment length. The full score calculation is shown in Equation 1 below:

$$\mathrm{MQM} = 100 \cdot (1 - \frac{10 \cdot \#\mathit{critical} + 5 \cdot \#\mathit{major} + \#\mathit{minor}}{\#\mathit{tokens}}) \quad (1)$$

The same type of MQM annotations were previously used in the WMT QE shared tasks for the document-level subtask (Fonseca et al., 2019; Specia et al., 2020a) also sponsored by Unbabel.

Severity	Category	Weight
Critical	all	10
Major	all	5
Minor	all	1

Table 3: Unbabel's MQM error weighting.

3.3 Human Evaluation Results

As discussed in Section 1, we decided to run our own human evaluation in order to generate our golden-truth ratings and come to stronger conclusions about the quality of each automatic metric across two domains. Unfortunately, this also meant that we were only able to evaluate a subset of documents of newstest2021 and TED talks. In Table 4, you can see the number of segments for each language pair and test set that we used for human evaluation.

language	news	TED
en→de	527/1002	529/606
en→ru	527/1003	512/684
zh→en	650/1948	529/843

Table 4: Numbers of MQM-annotated segments for each test set.

The results of the MQM human evaluation can be seen in Table 5. Most of the reference translations are ranked first, surpassing all MT systems, except for *ref-B* for zh→en TED talks and *ref-A* for en→de newstest2021. This confirms the findings in Freitag et al. (2021) that when human evaluation is conducted by professional translators and MQM, high-quality human translations typically still outperform MT. We will discuss the impact of the identified low-quality reference translations in Section 8.1.1 in more detail. We wish to highlight one more important observation: the ranking of the MT systems is sharply different when switching from the commonly used Newstest2021 test sets to TED talks. This is particularly interesting for the metrics task, as metrics need to assess MT quality purely on the basis of the translations themselves and cannot rely on features that are specific to any particular MT system. We will analyse the differences between Google's and Unbabel's MQM approach in Section 8.2 and compare our MQM human evaluation with the DA assessment from WMT in more detail in Section 8.3.

4 Metric Submissions and Baselines

4.1 Baselines

SacreBLEU baselines We use the following metrics from the SacreBLEU *v1.5.0* (Post, 2018) as baselines, with the default parameters:

- BLEU (Papineni et al., 2002) is the precision of n-grams of the MT output compared to the reference, weighted by a brevity penalty to Using SacreBLEU we obtained sentence-BLEU values using the `sentence_bleu` python function and for corpus-level BLEU we used `corpus_bleu`. Both functions were used with the default arguments.[6]

[6] BLEU+case.mixed+lang.LANGPAIR-+numrefs.1 +smooth.exp+tok.13a-+version.1.5.0

English→German ↓

System	news	TED
ref.C	0.48 (1)	n/a
ref.D	0.52 (2)	n/a
ref.B	0.80 (3)	n/a
VolcTrans-GLAT	1.04 (4)	1.49 (6)
Facebook-AI	1.05 (5)	1.06 (2)
ref.A	1.22 (6)	0.91 (1)
Nemo	1.34 (7)	2.14 (14)
HuaweiTSC	1.38 (8)	1.50 (7)
Online-W	1.46 (9)	1.12 (3)
UEdin	1.51 (10)	1.77 (11)
eTranslation	1.69 (11)	1.96 (13)
VolcTrans-AT	1.74 (12)	1.24 (4)
metricsystem4	2.05 (13)	1.78 (12)
metricsystem1	2.07 (14)	1.63 (8)
metricsystem3	2.27 (15)	1.44 (5)
metricsystem2	2.58 (16)	1.69 (9)
metricsystem5	2.61 (17)	1.72 (10)

Chinese→English ↓

System	news	TED
ref.B	4.271 (1)	0.42 (1)
ref.A	4.35 (2)	5.52 (15)
metricsystem1	4.42 (3)	1.90 (4)
metricsystem4	4.62 (4)	2.05 (7)
NiuTrans	4.63 (5)	2.49 (11)
SMU	4.84 (6)	2.202 (9)
MiSS	4.93 (7)	1.97 (5)
Borderline	4.94 (8)	2.40 (10)
metricsystem2	5.04 (9)	1.76 (3)
DIDI-NLP	5.09 (10)	1.65 (2)
IIE-MT	5.14 (11)	1.98 (6)
Facebook-AI	5.21 (12)	2.64 (12)
metricsystem3	5.39 (13)	2.99 (14)
Online-W	5.57 (14)	2.93 (13)
metricsystem5	6.39 (15)	2.15 (8)

English→Russian ↑

System	news	TED
ref-A	99.65 (1)	97.51 (1)
ref-B	98.40 (2)	n/a
Facebook-AI	92.75 (3)	87.40 (3)
Online-W	91.80 (4)	90.84 (2)
metricsystem4	91.25 (5)	70.63 (11)
metricsystem5	90.88 (6)	74.15 (7)
metricsystem1	90.79 (7)	72.08 (9)
metricsystem2	89.86 (8)	75.19 (6)
Online-A	87.87 (9)	71.93 (10)
Nemo	87.50 (10)	73.77 (8)
Online-G	87.23 (11)	77.62 (5)
Manifold	86.86 (12)	68.27 (13)
Online-B	85.66 (13)	78.05 (4)
metricsystem3	85.65 (14)	60.17 (15)
NiuTrans	83.47 (15)	69.94 (12)
Online-Y	79.27 (16)	61.91 (14)

Table 5: MQM human evaluations for Newstest2021 and TED. Lower average error counts represent higher MT quality for En→De and Zh→En (using Google's formulation of MQM), while higher scores represent higher quality for En→Ru (using Unbabel's MQM definition).

- TER (Snover et al., 2006) measures the number of edits (insertions, deletions, shifts and substitutions) required to transform the MT output to the reference. As in BLEU, for TER we used SacreBLEU `sentence_ter` and `corpus_ter` functions (with default arguments[7]) to obtain segment-level and system-level scores.

- CHRF (Popović, 2015) uses character n-grams instead of word n-grams to compare the MT output with the reference. For CHRF we used the SacreBLEU `sentence_chrf` function (with default arguments[8]) for segment-level scores and we average those scores to obtain a corpus-level score.

BERTscore BERTSCORE (Zhang et al., 2020) leverages contextual embeddings from pre-trained transformers to create soft-alignments between words in candidate and reference sentences using a cosine similarity. Based on the alignment matrix, BERTSCORE returns a precision, recall and F1 score. We used F1 without TF-IDF weighting.

Prism PRISM (Thompson and Post, 2020) is an automatic MT metric which uses a sequence-to-sequence paraphraser to score MT system outputs conditioned on their respective human references. We used the default parameters with version 0.1 and model *m39v1*.

4.2 Submissions

The rest of this section summarizes participating metrics.

COMET All COMET* metrics (Rei et al., 2021) were built using the Estimator architecture presented in Rei et al. (2020a,b). The difference between all the submitted metrics stem from: the data used for training, the size of the encoder model and whether or not they take advantage of the reference translation.

- COMET-DA_2020 is the same model submitted for last year's shared task (Rei et al., 2020b; Mathur et al., 2020b) while COMET-DA_2021 is a retrained version of the previous model that includes the DA udgements collected in 2020.

- COMET-MQM_2021 is an MQM adaptation of the COMET-DA_2021 model that further trains for 1 additional epoch on MQM z-scores extracted from the MQM ratings for newstest2020 provided for the task this year.

- COMETINHO-MQM and COMETINHO-DA are lightweight versions of COMET-MQM_2021 and COMET-DA_2021 respectively, which use a distilled version of XLM-RoBERTa as the encoder.

[7] TER+lang.LANGPAIR+tok.tercom-nonorm-punct noasian-uncased+version.1.5.0

[8] chrF2+lang.LANGPAIR- +numchars.6+space.false- +version.1.5.0.

	Metrics	broad category	Citation	Availability
Baselines	sentBLEU	lexical overlap	Papineni et al. (2002)	https://github.com/mjpost/sacrebleu
	BLEU	lexical overlap	Papineni et al. (2002)	https://github.com/mjpost/sacrebleu
	TER	lexical overlap	Snover et al. (2006)	https://github.com/mjpost/sacrebleu
	chrF	lexical overlap	Popović (2015)	https://github.com/mjpost/sacrebleu
	BERTScore	embedding similarity	Zhang et al. (2020)	https://github.com/Tiiiger/bert_score
	PRISM	MT-model-based	Thompson and Post (2020)	https://github.com/thompsonb/prism
Participants	COMET-*	neural finetuned metrics	Rei et al. (2021)	https://github.com/Unbabel/COMET
	OpenKiwi-MQM	neural finetuned metrics	Kepler et al. (2019)	https://github.com/Unbabel/OpenKiwi
	YiSi-*	embedding similarity	Lo (2019)	https://github.com/nrc-cnrc/yisi
	MTEQA	question-answer	Krubiński et al. (2021a)	https://github.com/ufal/MTEQA
	RegEMT-*	Ensemble	Stefanik et al. (2021)	https://github.com/MIR-MU/regemt
	RoBLEURT	neural finetuned metrics	Wan et al. (2021)	Not a public metric
	BLEURT-*	neural finetuned metrics	Sellam et al. (2020)	https://github.com/google-research/bleurt
	cushLEPOR-*	lexical overlap	Han et al. (2021)	https://github.com/poethan/cushLEPOR
	C-SPEC-*	neural finetuned metrics	Takahashi et al. (2021)	Not a public metric
	MEE-*	lexical and embedding similarity	Mukherjee et al. (2020)	https://github.com/AnanyaCoder/MEE_WMT2021

Table 6: Baseline metrics and participants of WMT21 Metrics Shared Task.

- Finally, COMET-QE-MQM_2021 and COMET-QE-DA_2021 are the reference-free versions of COMET-MQM_2021 and COMET-DA_2021 respectively.

From all the submitted models, the authors identified COMET-MQM_2021 and COMET-QE-MQM_2021 as their primary submissions to this years shared task edition.

OPENKIWI-MQM OPENKIWI-MQM (Kepler et al., 2019; Rei et al., 2021) is a multitask model that estimates a sentence-level MQM score along with word-level OK/BAD tags. The model is trained on top of XLM-RoBERTa using proprietary MQM data from several customer support domains. While word-level QE typically tags each word with an OK/BAD tag depending on post-edition information (Specia et al., 2020a), the OK/BAD tags used in OPENKIWI-MQM are derived directly from MQM annotation spans ignoring error types and/or severities.

YISI YISI (Lo, 2019) is a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources.

- YISI-1 is a reference-based MT evaluation metric. It measures the semantic similarity between a machine translation and human references by aggregating the idf-weighted lexical semantic similarities based on the contextual embeddings extracted from pretrained language models (e.g. BERT, CamemBERT, RoBERTa, XLM, XLM-RoBERTa, etc.).

- YISI-2 is the bilingual, reference-less version for MT quality estimation. It uses bilingual mappings of the contextual embeddings extracted from multilingual pretrained language models (e.g. XLM-RoBERTa) to evaluate the crosslingual lexical semantic similarity between the input and MT output.

YISI is an untrained metric and the submissions this year are the same as those in WMT20. The metric settings are described in Lo (2020) and Lo and Larkin (2020).

MTEQA MTEQA (Krubiński et al., 2021a,b) is an MT evaluation metric that leverages automatically generated questions and answers to assess the quality of MT systems. It builds upon the assumption that a good translation should preserve all of the key information that one can extract from the reference. Based on syntactic structure and NER system, they extract potential answers from the reference, and for each of them generate a human readable question. They then use a question-answering system to provide a new (test) answer given the question and the MT output as the context. The test answer is then compared to the reference answer to obtain the numerical score.

REGEMT REGEMT (Stefanik et al., 2021) is a family of ensemble metrics trained on MQM labels.

- {SRC, TGT}-REGEMT: This ensemble combine selected metrics of surface-, syntactic- and semantic-level similarity as input features to a regression model that estimates a quality assessment. Some of these features are newly introduced and some are based on related work. The reference-free ensemble uses as input features: Source length, Target length, Contextual SCM, Contextual WMD, BERTScore, Prism and Compositionality the reference-base ensemble uses: COMET, BLEURT, BLEU, METEOR, Non-contextual SCM and WMD.

- REGEMT-BASELINE: This ensemble uses only Source length and Target length of the given texts, in characters

The authors identified {SRC, TGT}-REGEMT as their primary submissions.

ROBLEURT ROBLEURT (Wan et al., 2021), short for Robustly Optimizing the training of BLEURT, is a model-based metric based on powerful language model XLM-RoBERTa. The ROBLEURT metric is constructed by the following steps: 1) jointly leveraging the advantages of source-included and reference-only metric models, 2) continuously pre-training the model with massive synthetic data produced by the real-world machine translation engines, and 3) fine-tuning the model with a data denoising strategy.

BLEURT BLEURT-20 and BLEURT-21-BETA are obtained by fine-tuning Rebalanced mBERT (Chung et al., 2021) (a multilingual variant of BERT) on a combination of two datasets: previous ratings from the WMT shared, task and generated data. The generated data consists of “perfect" sentence pairs, obtained by copying the reference into the hypothesis, as well as “catastrophic" sentence pairs, obtained by randomly sampling tokens for each language pair. The fine-tuning methodology is similar to (Sellam et al., 2020). BLEURT-20 was trained on human ratings from WMT metrics 2015 to 2019 (z-scores) using WMT20 for test, and BLEURT-21-BETA was trained on WMT 2015 to 2020. The suffixes “-20” and “-21” denote the year of the WMT Metrics ratings that were used to build the test sets. The authors identified BLEURT-20 as their primary submission.

hLEPOR and cushLEPOR

- HLEPOR (Han et al., 2013) is an augmented metric with factors including enhanced sentence length penalty, precision, recall, and positional difference penalty which captures word order.

- CUSHLEPOR(LM) (Han et al., 2021) is a customized hLEPOR metric that uses LABSE pre-trained language model to automatically optimise hLEPOR parameters towards better correlation to human judgement and lower cost.

- CUSHLEPOR(PSQM) (Han et al., 2021) is trained and validated on the MQM and pSQM annotations from human professionals (Freitag et al., 2021). The tuned cushLEPOR achieves very high agreement to LABSE pre-trained language model in performance but uses much less computational cost as a distilled model.

The authors identified CUSHLEPOR(LM) as their primary submission.

C-SPEC C-SPEC (Takahashi et al., 2021) is designed for both segment-level and system-level translation evaluation. The authors’ objective was to design a better metric by detecting significant translation errors that would not be ignored in real instances of human evaluation. Thus, pseudo-negative examples are generated in which selected words in the translation are replaced with alternatives based on a Word Attribute Transfer, and a metric model is built to handle such serious translation errors (denoted as C-SPECPN). A multilingual large pretrained model is fine-tuned on the provided corpus of past years’ metrics task and fine-tuned again further on the synthetic negative samples that is derived from the same fine-tuned corpus. The authors identified C-SPECPN as their primary submission.

MEE

- MEE (Mukherjee et al., 2020) is an automatic evaluation metric that leverages the similarity between embeddings of words in candidate and reference sentences to assess translation quality focusing mainly on adequacy. Unigrams are matched based on their surface forms, root forms and meanings which aids to capture lexical, morphological and semantic equivalence. Semantic evaluation is achieved by using pretrained fasttext embeddings provided by Facebook to calculate the word similarity score between the candidate and the reference words. MEE computes evaluation scores using three modules namely exact match, root match and synonym match. In each module, fmean-score is calculated using harmonic mean of precision and recall by assigning more weight to recall. A final translation score is obtained by taking the average of fmean-scores from the individual modules.

- MEE2 is an improved version of MEE, focusing on computing contextual and syntactic equivalences along with lexical, morphological and semantic similarity. The intent of MEE2 is to capture fluency and context of the MT outputs along with their adequacy. Fluency is captured using syntactic similarity and context is captured using sentence similarity leveraging sentence embeddings. The final sentence translation score is the weighted combination of three similarity scores: a) Syntactic Similarity achieved by modified BLEU score; b) Lexical, Morphological and Semantic Similarity: measured by explicit unigram matching similar to MEE score; c) Contextual Similarity: Sentence similarity scores are calculated by leveraging sentence embeddings of *Language-Agnostic BERT* models.

The authors identified MEE2 as their primary submission.

5 Main Results

Currently, the main use case of automatic metrics is to rank systems either during system development or by comparing your own output with the one from other research institutes or competitors. Consequently, we present system-level correlations as our main metric in this year's WMT21 metrics task. To be in line with the main use case, we present *pairwise accuracy* numbers for each metric that calculate the accuracy scores on binary comparison of system outputs for each language pair. We refer the reader to Section 7 for language pair specific results on both the segment and system level with more traditional correlation metrics.

5.1 System-Level

The system-level metric scores submitted by the participants pertained to the complete WMT test set, but we collected human MQM scores for only a subset of documents, as shown in Table 4. To correct for this discrepancy, we re-computed system-level scores as averages over the segments for which MQM scores were available, after first verifying with all participants that their system-level scores were computed in the same fashion.

To generate a single score combining the data from all 3 language pairs, we calculate *pairwise accuracy* (Kocmi et al., 2021) as our primary scoring metric. Pairwise accuracy is defined as follows: For each language pair and system pair, we calculate the difference of the metric scores (metricΔ) and the difference in average human judgements (humanΔ) for each system pair. We calculate accuracy for a given metric as the number of rank agreements between the metric and human deltas, divided by the total number of comparisons:

$$\text{Pairwise accuracy} = \frac{|\text{sign}(\text{metric}\Delta) \;=\; \text{sign}(\text{human}\Delta)|}{|\text{all system pairs}|} \tag{2}$$

We present results for three different settings: Looking at the news domain with and without human translations (HT) as additional systems: (a) *Newstest2021 w/o HT*, (b) *Newstest2021 w/ HT*, and (c) looking at *TED talks w/o HT*. In this section, we consider only the primary submissions of each metric team and the baseline metrics. We have multiple reference translations for some settings. Instead of reporting results with respect to all reference translations, we use here for reference-based metrics only the single reference that was judged best by the MQM raters for each language pair. The remaining reference translations are used in the role of participating MT systems in the "w/ HT" evaluations. Table 7 summarizes the use of reference translations for different language pairs and domains. We will analyse the impact of using different reference translations in Section 8.1 in more detail.

language	news		TED
	best ref	scored refs	best ref
en→de	C	A, D	A
en→ru	A	B	A
zh→en	B	A	B

Table 7: Use of reference translations.

Metric rankings based on pairwise accuracy can be found in Table 8. The top significance cluster (bolded in the table) consists of primary or baseline metrics that are not significantly outperformed by any other primary or baseline metrics nor outperformed by a primary or baseline metric not in the top cluster.[9]

Metric	newstest21 w/o HT	newstest21 w/ HT	TED w/o HT
tgt-regEMT	**0.773** (1)	0.694 (5)	0.636 (15)
Prism	**0.769** (2)	0.641 (7)	0.733 (5)
cushLEPOR(LM)	**0.763** (3)	0.622 (9)	0.647 (14)
C-SPECpn	**0.757** (4)	**0.784** (1)	0.704 (10)
bleurt-20	**0.753** (5)	0.718 (3)	0.749 (3)
MEE2	**0.753** (6)	0.628 (8)	0.713 (7)
BERTScore	**0.745** (7)	0.621 (10)	0.721 (6)
chrF	**0.745** (8)	0.621 (11)	0.713 (8)
BLEU	0.741 (9)	0.618 (12)	0.741 (4)
YiSi-1	0.737 (10)	0.615 (13)	**0.757** (2)
COMET-QE-MQM_21	0.733 (11)	**0.774** (2)	0.652 (13)
COMET-MQM_21	0.713 (12)	0.688 (6)	**0.773** (1)
MTEQA*	0.705 (13)	0.577 (15)	0.705 (9)
TER	0.696 (14)	0.585 (14)	0.636 (16)
OpenKiwi-MQM	0.692 (15)	0.698 (4)	0.680 (12)
RoBLEURT*	0.641 (16)	0.549 (16)	0.692 (11)
YiSi-2	0.510 (17)	0.429 (17)	0.494 (17)
src-regEMT	0.494 (18)	0.415 (18)	0.405 (18)

Table 8: Pairwise accuracy for Chinese→English, English→German, and English→Russian using the MQM annotations. Correlations for metrics in the top significance cluster are bolded. All submissions labelled with * participated only in 1 or 2 language pairs and are not considered during significance testing. Metrics not using reference translation (QE-metrics) are indicated by italics.

- **Newstest2021 w/o HT** This setting is most similar to previous years' settings. Metrics are required to score all MT outputs without considering human translations (HT). This setup investigates how metrics evaluate current SOTA MT

[9] Note that this definition is different from the metric clustering used in previous metrics tasks, in which every metric in a cluster must be significantly better than all metrics in lower clusters.

models. Looking at the ranking in Table 8, we can see that in total 8 metrics fall into the first significance cluster. The cluster includes a variety of embedding-based metrics and surface metrics. None of the QE metrics (i.e. reference-less metrics) are part of the first cluster.

- **Newstest2021 w/ HT** When considering the additional reference translations as system outputs (ref-A for zh→en, ref-B for en→ru, ref-A and ref-D for en→de), the ranking of the metrics is sharply revised. The QE metric *COMET-QE-MQM_2021* and the reference-based metric *C-SPECpn* are the winners in this setup. Overall, the embedding-based metrics that also rely on fine-tuning are much better in rating human translation higher than MT output and thus dominate this setting.

- **TED talks w/o HT** This year, we also measured the domain robustness of each metric on the TED talks domain. In Table 8, we can see that *COMET-MQM_2021* and *YiSi-1* show the highest correlation with human ratings on the TED domain. Interestingly, both metrics did not fall into the first significance cluster in the previous two settings of the news domain, leading to very different conclusion about the quality of metrics.

5.2 Significance Testing

We run PERM-BOTH hypothesis test (Deutsch et al., 2021) on the pairwise system-level accuracy of Table 8 to measure significance between metrics' performance.[10] Results can be seen in Figure 1. By looking at the heat map of Newstest2021 without human translations (*Newstest2021 w/o HT*), we observe that the top performing metrics are not significantly different. This observation changes when we add human translations to the setup (*Newstest2021 w/ HT*). The top 2 performing metrics, although different ones, are significantly better than all other metrics. This setup gives us the clearest result of all our 3 different setups and highlights that embedding-based metrics that are fine-tuned on previous years' human ratings rate human translations much better than all the other metrics and are good at distinguishing human-produced text. Another different situation can be seen when looking at the TED talk setting (*TED talks w/o HT*). Even though we see more significant differences compared to *Newstest2021 w/o HT*, most pairs of metrics are not significantly different.

6 Challenge Sets

While the correlation analysis is testing the evaluation metrics on their ability to rank MT systems according to translation quality, we are also interested in understanding metrics' performance on identifying certain types of translation errors. We created three challenge sets containing translation errors that are believed to be challenging for automatic MT evaluation metrics to identify. A good metric should not only rank candidate translations by their quality but also be sufficiently sensitive to these types of errors.

Each challenge set consists of two MT outputs (and the corresponding source and reference) where one of them contains the type of translation error of interest and the other does not. Metrics are expected to give a lower score to the MT output containing the error.

We use Kendall's tau-like correlation, typically used for DARR (Bojar et al., 2017; Ma et al., 2018, 2019; Mathur et al., 2020b), for evaluating the primary submissions on the challenge sets. Kendall's tau-like correlation is defined as follows:

$$\tau = \frac{\mathit{Concordant} - \mathit{Discordant}}{\mathit{Concordant} + \mathit{Discordant}} \tag{3}$$

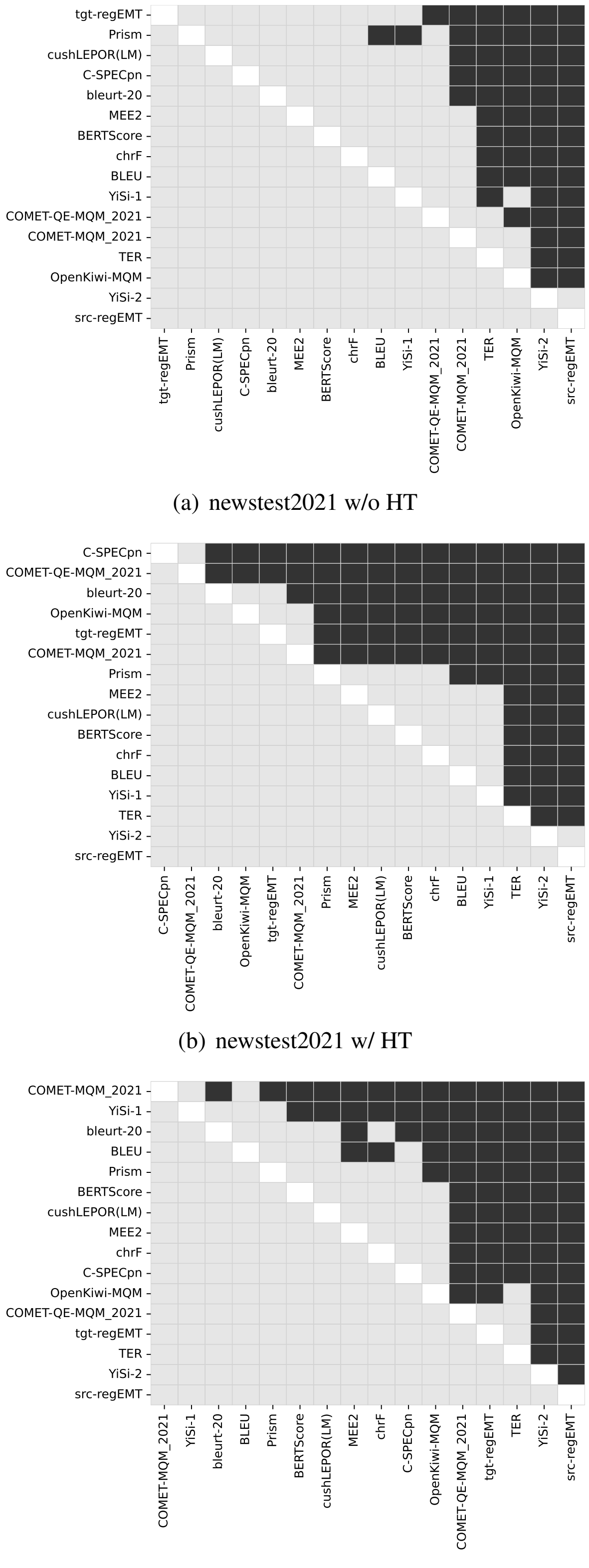
where *Concordant* is the number of times a metric assigns a higher score to the MT output without the error and *Discordant* is the number of times a metric assigns a higher score to the MT output containing the error of interest.

6.1 Negation and Sentiment Polarity Challenge Set

The goal of this challenge set is to test metrics' ability to penalize translations when there is a catastrophic error in reversing of a negation or of sentiment polarity. It is a common phenomenon that MT systems may either introduce or remove a negation (with or without an explicit negation word), or may reverse the sentiment polarity of the sentence (e.g. a negative sentence becomes positive or vice-versa). These types of errors could result in serious consequences of misleading users of MT.

The WMT2020 MT Robustness shared task (Specia et al., 2020b) collected Wikipedia Edit comments with toxic content that could lead to possible

[10] Previous editions of the metrics task used the Williams test (Williams, 1959), but we adopted PERM-BOTH because it is more general, and because Deutsch et al (2021) demonstrate that it has higher power.



(a) newstest2021 w/o HT

(b) newstest2021 w/ HT

(c) TED talks w/o HT

Figure 1: The results of running PERM-BOTH hypothesis test to find a significant difference between metrics' pairwise system-level accuracy. Dark squares mean the row metric correlates significantly better than the column metric at $\alpha = 0.05$.

catastrophic errors in the MT output. After selecting segments of interest they created reference translations for the entire test set using professional translators. Finally, they collected annotations of catastrophic errors on the translations performed by participating systems[11].

To test metrics on sentiment polarity we looked for source sentences from the English→German data portion where we can find an MT output annotated with a sentiment polarity error and another MT output without the polarity error. The resulting challenge set contains 177 source sentences (not necessarily distinct), each equipped with two MT outputs, one with a catastrophic error and one without it. We note that most of the sentences in this challenge set contain toxic language.

Table 9 shows the results for this challenge set. We also show the actual number of concordant pairs here because this challenge set is rather small. Despite the high severity of the translation error in reversing the sentiment polarity or negation, we see that both the baselines and the submissions struggle to accurately discriminate between translations with and without such errors. TER and BERTSCORE are the only two metrics that are able to achieve a medium correlation (i.e. greater than 0.4) with human annotators on ranking the translation with the catastrophic error as lower in translation quality. Perhaps more importantly, embedding-based and semantic-oriented metrics, such as BERTSCORE, YISI-1, etc., do not significantly outperform surface-form matching metrics, such as TER, CHRF and SENT-BLEU. This may indicate that the pretrained language models used by the embedding-based metrics are weak at learning language representations that explicitly reflect differences in negation and sentiment polarity.

6.2 Corrupted Reference Challenge Set

The goal of this challenge set is to sanity check the behaviour of the submitted metrics and possibly identify some weaknesses in detecting specific anomalies in a corrupted reference translation. In order to do this we used this years' Chinese→English Newstest corpus, which contains two human systems (referenceA and referenceB) and we perturb one of these human systems while using the other as reference. Given that, our final corpus is composed of $14,080$ tuples with

[11] Professional translators with access to the original source sentence, the reference and the system output were used during this evaluation.

Metric	*Concordant*	τ
TER	132	0.492
BERTSCORE	124	0.401
CHRF	123	0.390
YISI-1	122	0.379
MEE2	120	0.356
BLEURT-20	119	0.345
SENT-BLEU	118	0.333
C-SPECPN	118	0.333
COMET-MQM_2021	117	0.322
TGT-REGEMT	115	0.299
SRC-REGEMT	112	0.266
CUSHLEPOR(LM)	108	0.220
OPENKIWI-MQM	108	0.220
PRISM	107	0.209
MTEQA	106	0.198
COMET-QE-MQM_2021	106	0.198
YISI-2	104	0.175

Table 9: Results for the Negation and Sentiment Polarity Challenge Set. (Out of 177 hypothesis pairs)

(source, referenceBpert, referenceB, referenceA) where referenceBpert denotes the perturbed reference.

The perturbations used are: *antonym replacement*, *word omission*, *tokenization*, *sentence omission*,[12] *punctuation removal*, *number swapping*, *lowercasing*, *word addition* and addition of *spelling errors*. Table 22 in Appendix A shows examples for each perturbation.

From Table 10 we can observe that for most embedding based metrics (YISI, BERTSCORE, BLEURT-21-BETA, ROBLEURT, PRISM) correlations are close to 1.0 for all perturbation types. The only exceptions are COMET-MQM_2021 and C-SPECPN that seem to struggle with *sentence omission* and *punctuation removal*. This behaviour is even more unexpected if we take into consideration that they seem to be sensitive to *word omission*. Regarding punctuation removal, since both metrics are fine-tuned on Google MQM annotations (see Section 3.1) we hypothesize that they learn to be less sensitive to punctuation errors. Regarding the lexical metrics, we can observe that SENT-BLEU, CHRF and CUSHLEPOR(LM) are not sensitive to tokenized text. This is an expected behaviour for lexical metrics since they typically ignore whitespaces. Also, CUSHLEPOR(LM) scores -1.0 in lowercased text. This seems to indicate that this metric does not encode casing information.

6.3 German→English Challenge Set

The challenge set is based on the test suite by Macketanz et al. (2018a). It is a test suite for German-English that consists of around 5,500 German test sentences covering 107 linguistically motivated phenomena (listed in Avramidis et al. (2020)), organized in 14 categories. These phenomena do not follow a linguistic theory but rather cover various grammatical aspects which are relevant for MT. Each phenomenon is represented by at least 20 test sentences to guarantee a balanced test set. The test suite is used to evaluate MT systems with regard to their performance on the test sentences. The evaluation operates semi-automatically and is based on a set of handwritten rules which contain regular expressions and fixed strings.

The test suite has been used to evaluate the outputs of 40 German-English systems submitted at the translation task of the Conference of Machine Translation (WMT) for three consecutive years (Macketanz et al., 2018b; Avramidis et al., 2019, 2020) and also this year (Macketanz et al., 2021). Across the past three years, this amounts to 40 system outputs. We use these outputs to construct the challenge items for the metrics task, since the test suite contains only source sentences and handwritten rules for the outputs but no reference translations. For every source sentence of the test suite we separate MT outputs into “correct” and “incorrect” ones using the handwritten rules of the test suite and create a tuple including; (1) a set of “correct” MT outputs, to be given to the metrics as supposedly correct reference translations and, (2) a pair of one “incorrect” and one “correct” translation randomly sampled from the respective set. Note that the “correct” candidate does appear among the references (1). The goal of the metric is to score the “incorrect” translation worse than the “correct” one.

The same source sentence may be appear more than once, if there is more than one WMT translation marked as wrong by the rules for this item. The above process resulted in a metrics challenge set with 1,819 items with source, wrong hypothesis, correct hypothesis, and a pseudo-reference (another MT that was deemed correct for that phenomenon).

[12] Note that after experimenting with paragraph-level translation in WMT20, WMT21 moved back to segments again corresponding to individual sentences. In Chinese→English corpus, paragraph boundaries are not apparent (all documents consist of one paragraph). For the purposes of this experiment, we used *nltk.sentence_tokenizer* and looked for all the references B with more than 1 sentence and randomly delete 1 of those sentences to create referenceBpert. Note that since we do not have entire paragraph, the size of this challenge is 88 samples only.

Metric	Antonym	W. Omission	Tokenized	Sent. Omission	Punct.	Numbers	Lower.	W. Add.	Spell.
SENT-BLEU	0.792	0.787	-0.617	0.409	0.640	0.715	0.633	0.986	0.954
TER	0.994	0.597	0.966	0.568	0.739	0.996	1.000	1.000	0.997
CHRF	0.887	0.983	-0.516	0.523	0.761	0.899	0.708	0.903	0.981
BERTSCORE	0.986	0.984	0.994	0.909	0.950	0.993	0.799	0.996	0.998
PRISM	0.998	0.995	0.972	0.864	0.990	1.000	0.969	1.000	0.999
MTEQA	0.329	0.721	-0.522	0.273	-0.340	0.712	-0.415	0.649	0.624
YISI-1	0.991	0.996	0.993	0.977	0.951	0.996	0.960	0.999	1.000
BLEURT-20	0.992	0.989	0.983	0.909	0.931	0.993	0.976	0.998	0.997
COMET-MQM_2021	0.996	0.994	0.994	-0.068	0.235	0.993	0.965	0.993	1.000
C-SPECPN	0.991	0.988	0.576	0.409	0.622	0.876	0.922	0.991	0.996
CUSHLEPOR(LM)	0.826	0.779	-0.431	0.500	0.877	0.730	-1.000	0.982	0.957
MEE2	0.975	0.968	0.681	0.955	0.853	0.981	0.855	0.987	0.989
ROBLEURT	0.998	0.991	0.995	0.818	0.919	1.000	0.986	0.996	1.000
TGT-REGEMT	0.930	0.772	0.675	0.364	0.599	0.798	0.510	0.923	0.978
YISI-2	0.979	0.953	0.542	0.977	0.947	0.835	0.806	0.991	0.990
COMET-QE-MQM_2021	0.991	0.983	0.983	-0.318	-0.199	0.989	0.931	0.982	0.998
OPENKIWI-MQM	0.962	0.952	0.070	0.091	0.797	0.243	0.719	0.979	0.991
SRC-REGEMT	0.637	0.512	0.357	0.341	0.209	0.333	0.365	0.342	0.300

Table 10: Kendall's tau-like correlation results for the Corrupted References Challenge set (Section 6.2). The horizontal lines delimit baseline metrics (top), participating reference-based metrics (middle) and participating reference-free metrics (bottom).

The covered phenomena are: *Function Words (FW), Non-verbal Agreement (NVA), Verb Tense/Aspect/Mood (VT), Composition (Comp.), Multi-Word Expressions Negation (MWE N.), Punctuation (Punct.), Verb Valency (VV), Subordination (Sub.), Coordination and Ellipsis (CE), Named Entities and Terminology and Long Distance Dependencies and Interrogative (LDD).*

Overall, from Table 11 we observe that embedding-based metrics such as BLEURT-20 and COMET-MQM_2021 seem to be less sensitive to *Subordination*, *Named Entities and Terminology*, and to *Punctuation*. We can also observe a clear performance difference between reference-free and reference-based metrics. Nonetheless most metrics have positive correlations in all covered phenomena. Note that this corpus is composed of "pseudo-references" which can have a negative impact on metrics' performance (see Section 8.1).

7 Results per Language Pair

We computed individual correlation results for each focus language pair (English→German, English→Russian, Chinese→English) at both the system and segment level. The system-level metric scores submitted by the participants pertained to the complete WMT test set, but we collected human MQM scores for only a subset of documents, as shown in Table 4. To correct for this discrepancy, we re-computed system-level scores as averages over the segments for which MQM scores were available, after first verifying with all participants that their system-level scores were computed in the same fashion.[13] Exceptions to this pattern are the baseline metrics BLEU and TER: the system-level versions of these metrics are not averages over segment-level scores, and we computed them only over the MQM segments.

Since we have multiple reference translations for the focus language pairs, we required participants to submit versions of their (reference-based) metrics for each reference. We used only the scores corresponding to the reference that was judged best by the MQM raters for each language pair. For the news domain, we evaluated metric performance both when using only MT outputs and using MT outputs augmented by human references, adding all remaining references in the latter condition except for English→German, where we excluded reference B since it was very similar to the best reference C. Table 7 summarizes the use of reference translations for different language pairs and domains.

We measure correlation using the Pearson-r statistic at the system level and the Kendall-tau statistic at the segment level. Pearson correlation is complementary to the pairwise accuracy used for our global results as discussed in Section 5: it tests linear fit with MQM scores, a stringent but

[13] In contrast to the standard practice with WMT DA scores, where scored segments for each system are sampled independently, the segments for which we have MQM scores comprise a fixed set, independent of the MT system being scored.

Metric	FW	NVA	FF	VT	Comp.	Amb.	MWE N.	Neg.	Punct.	VV	Sub.	CE	NE & Term.	LDD
Sent-BLEU	0.50	0.66	0.32	0.37	0.09	0.42	0.38	0.77	0.64	0.42	0.48	0.43	0.30	0.52
TER	0.64	0.80	0.74	0.67	0.43	0.60	0.58	0.76	0.71	0.59	0.59	0.6	0.53	0.65
chrF	0.42	0.56	0.63	0.57	0.37	0.70	0.46	0.71	0.43	0.51	0.38	0.45	0.44	0.54
BERTScore	0.61	0.59	0.89	0.76	0.76	0.74	0.60	0.71	0.68	0.77	0.47	0.63	0.48	0.68
Prism	0.72	0.56	0.74	0.82	0.74	0.82	0.70	0.65	0.58	0.80	0.45	0.71	0.53	0.71
MTEQA	-0.64	-0.38	0.26	-0.77	0.03	0.54	-0.05	-0.59	-0.87	-0.57	-0.58	-0.30	0.34	-0.34
YiSi-1	0.63	0.58	0.95	0.80	0.76	0.77	0.64	0.76	0.62	0.82	0.40	0.61	0.60	0.68
BLEURT-20	0.70	0.58	0.79	0.72	0.68	0.83	0.65	0.65	0.30	0.72	0.49	0.68	0.38	0.73
COMET-MQM	0.58	0.55	0.89	0.76	0.52	0.74	0.66	0.71	0.33	0.74	0.41	0.67	0.44	0.67
C-SPECpn	0.45	0.45	0.47	0.56	0.35	0.83	0.54	0.41	0.24	0.57	0.33	0.65	0.38	0.58
RoBLEURT	0.60	0.65	0.68	0.77	0.64	0.77	0.68	0.71	0.30	0.81	0.39	0.58	0.38	0.70
tgt-regEMT	0.54	0.53	0.47	0.38	0.07	0.36	0.21	0.29	0.31	0.32	0.18	0.23	0.38	0.35
YiSi-2	0.10	-0.03	0.11	0.36	0.37	0.29	0.03	0.18	0.45	0.35	0.11	0.25	0.13	0.42
COMET-QE-MQM	0.47	0.41	0.63	0.27	0.33	0.52	0.37	0.53	0.09	0.53	0.31	0.49	0.17	0.61
OpenKiwi-MQM	0.42	0.25	0.37	0.45	0.23	0.47	0.44	0.53	0.41	0.56	0.27	0.62	0.27	0.47
src-regEMT	0.45	0.08	-0.05	0.29	0.41	0.26	0.00	-0.06	0.37	0.44	0.05	0.05	0.12	0.18
Average	0.45	0.43	0.56	0.49	0.42	0.60	0.43	0.48	0.35	0.52	0.30	0.46	0.37	0.51

Table 11: Kendall's tau-like correlation results for the German→English challenge set based on (Macketanz et al., 2018a) test suite. Note that not all metrics submitted to this challenge set hence some metrics are missing. The horizontal lines delimit baseline metrics (top), participating reference-based metrics (middle) and participating reference-free metrics (bottom).

Metric	Total	Language Pair			Granularity		Data condition		
	"wins"	en→de	en→ru	zh→en	sys	seg	news w/o HT	news w/ HT	TED
C-SPECpn	11	4	3	4	6	5	3	5	3
bleurt-20	10	4	5	1	4	6	4	3	3
COMET-MQM_2021	10	3	3	4	3	7	3	2	5
tgt-regEMT	4	1	1	2	3	1	2	1	1
COMET-QE-MQM_2021	3	1	1	1	3			3	
OpenKiwi-MQM	3	2		1	3		1	2	
RoBLEURT*	3			3	1	2	1		2
cushLEPOR(LM)	2	1		1	2		1		1
BERTScore	2	1	1		2		1		1
Prism	2		2		2		1		1
YiSi-1	2		2		2		1		1
MEE2	2	2			2		1		1
BLEU	1	1			1		1		
hLEPOR	1		1		1				1
MTEQA*	1			1	1				1
TER	1			1	1				1
chrF	1			1	1				1

Table 12: Summary of language-specific results. Numbers give the count of times each primary metric occurred in the top cluster for the specified condition. Metrics not being among the winners in any competition are not listed. Reference-free metrics are indicated by italics. All submissions labelled with * participated only in 1 or 2 language pairs.

reasonable criterion since we expect these scores to conform to a linear scale (for example, a translation with two minor errors is twice as bad as one with only a single error). Pearson has well-known drawbacks (Mathur et al., 2020a), notably sensitivity to outliers, which we avoided by choosing only relatively high-performing systems. In preliminary tests, Pearson also yielded a larger number of pairwise significant differences among metrics than Kendall, an important property since our fairly small number of systems makes it difficult to reliably distinguish metrics at the system level.

Segment-level scores—metric or human—are naturally arranged as a system × segment matrix (rows × columns). There are several ways to extract vectors for input to correlation statistics. Comparing metric and human row vectors corresponds to a use case of judging the relative quality of different segments output by a given MT system ("where is my system making mistakes on this test set?"); comparing column vectors corresponds to judging the relative quality of outputs for a given source segment across different MT systems ("which systems performed better or worse than mine on this segment?”). To avoid emphasizing either of these scenarios at the expense of the other, we flattened the metric and human score matrices into single vectors (row1, row2, ...) before comparing them. This measures the metrics’ ability to assign independent scores to MT segments, abstracting away from system or source segment, and provides a large number of comparisons to boost statistical significance. We used Kendall rather than Pearson correlation for robustness to segment-level noise.[14]

The results for each language pair and granularity are shown in Tables 23 to 28, with corresponding pairwise significance plots derived using the PERM-BOTH test in Figures 2 to 7. The tables contain results for all metrics; the significance plots include only primary and baseline metrics. In the tables, primary submissions are in bold, baseline metrics are underlined, and metrics that used only the source have “-src” appended to their name. For each condition (news without human translations, news with human translations, or TED), the *scores* of primary and baseline metrics in the top cluster are in bold. The top cluster consists of primary or baseline metrics that are not significantly outperformed by other primary or baseline metrics nor outperformed by a primary or baseline metric not in the top cluster.[15] In the significance plots, this corresponds to the leftmost block of columns containing no dark squares.

Table 12 summarizes all results in this section by counting the number of times each metric occurs in the top cluster (it got a “win”), summed across different ways of partitioning the results. This synthesis is fairly crude, since it treats all conditions as equally important. Also, membership in the top cluster is likely to be subject to high statistical variance, and metrics that fall outside this cluster are not accounted for; in particular, those that sometimes perform very poorly are not penalized. Nevertheless, the counts permit some general observations.

In terms of total “wins”, three metrics stand out clearly: C-SPECPN, BLEURT-20, and COMET-MQM_2021. These have fairly evenly-distributed performance across languages, granularities, and data conditions, with the exception of BLEURT-20, which does relatively poorly on Chinese→English. Their advantage over other metrics is most pronounced at the segment level and when human translations are included among the systems to be judged (*w/ HT*)—both of which are more challenging tasks. In contrast, the distribution of metrics that achieve top-level performance is much broader for system-level granularity, the out-of-domain TED setting, and to a lesser extent the news *w/o HT* setting. Two metrics that do not use a reference translation—COMET-QE-MQM_2021-src and OpenKiwi-MQM-src—do surprisingly well overall, particularly in the *w/ HT* condition, but perform poorly at the segment level. This could be explained by these metrics benefiting from their ability to distinguish human vs. machine produced text. Finally, the surface-level baselines—BLEU, TER, and chrF—join the winners exclusively at the system level and almost exclusively in the out-of-domain TED condition.

8 Additional Results

8.1 Impact of Reference Translation

The quality of the reference translation can have a higher impact on the correlation to human ratings than the actual choice of metric (Freitag et al., 2020). For all our different test sets and language pairs, we consequently included all reference translations in our human evaluation to (a) assure that we have reference translation with high quality and (b) to choose the best reference translation for our main results. In this section, we present two interesting observations by looking into the Chinese→English TED talks and the English→German news setups.

8.1.1 zh→en TED

We started by having only one reference translation for all TED talks. Unfortunately, the MQM evaluation revealed that the reference translation *ref-A* for Chinese→English was ranked last – lower than all the MT systems – and that it contained on average more than one major error (= 5 MQM points) per segment. We spot checked the errors and agreed that the reference translation indeed contained many errors. We then decided to acquire

[14] Our use of Kendall differs in two major aspects from the “Kendall-like” statistic used for segment-level correlations in previous editions of the WMT metrics task: we do not threshold MQM score differences, as we consider them to be more reliable than DA scores; and we compare all pairs of scores over complete flattened matrices rather than comparing pairs of scores in each column, and micro-averaging results across columns.

[15] Note that this definition is different from the metric clustering used in previous metrics tasks, in which every metric in a cluster must be significantly better than all metrics in lower clusters.

	ref-A	ref-B
MQM	5.52	0.42
MTEQA	0.47 (3)	0.74 (1)
TER	0.40 (9)	0.71 (2)
BERTScore	0.42 (6)	0.69 (3)
bleurt-20	0.45 (5)	0.68 (4)
cushLEPOR (LM)	0.39 (11)	0.68 (5)
Prism	0.46 (4)	0.68 (6)
COMET-MQM_2021	0.40 (8)	0.67 (7)
BLEU	0.30 (13)	0.65 (8)
YiSi-1	0.42 (7)	0.65 (9)
chrF	0.40 (10)	0.62 (10)
MEE2	0.36 (12)	0.60 (11)
C-SPECpn	0.49 (2)	0.54 (12)
tgt-regEMT	0.5 (1)	0.37 (13)
average	0.42	0.64

Table 13: Pairwise accuracy for ranking system pairs for TED Chinese→English using either ref-A (original ref) or ref-B (extra generated ref).

a new reference translation (*ref-B*) which turned out to be better than all MT systems after running a human evaluation. The impact of using an excellent versus an inaccurate human reference translation can be seen in Table 13. All metrics achieve an accuracy score lower than 0.5 when using ref-A to calculate their scores. This means that the metrics would perform worse than by chance. By switching to ref-B, all but one metric (*tgt-regEMT*) greatly improve their correlation score. This demonstrates once again that metrics become unreliable when they are provided with inaccurate reference translations.

8.1.2 en→de Newstest2021

For English→German Newstest2021, we started with two reference translations (*ref-A* and *ref-B*). Both reference translations had issues: ref-A was ranked lower than two MT systems (see Table 5) and we agreed with that assessment after spot checking the errors. ref-B had high-levels of overlap with the Online-W MT system and is most likely a post-edited translation of Online-W. Therefore, Microsoft and Google sponsored two new reference translations (*ref-C* and *ref-D*) which turned out to be the best translations based on MQM. In Table 14, you can see the pairwise accuracy scores from all reference-based primary and baseline metrics. Despite the good (low) MQM scores of both ref-C and ref-D, the ranking of the metrics when using these two references is quite different. Some metrics are more robust when switching the reference translation (e.g. *Prism*, *YiSi-1*, or *C-SPECpn*, but others yield very different correlation scores (e.g. *BERTScore*, *tgt-refEMT*, or *BLEU*). Something else in addition to quality makes *ref-C* more appealing for metrics than *ref-D*. We do not have an explanation why the quality of some metrics is so different when switching the reference translation and leave this as an open challenge for the community to better understand why this is happening.

	ref-A	ref-C	ref-D
MQM	1.22	0.48	0.52
BERTScore	0.91 (1)	0.94 (1)	0.77 (10)
cushLEPOR (LM)	0.81 (10)	0.92 (2)	0.81 (6)
BLEU	0.87 (5)	0.90 (3)	0.69 (12)
MEE2	0.87 (4)	0.90 (4)	0.80 (8)
TER	0.89 (2)	0.90 (5)	0.80 (7)
chrF	0.82 (8)	0.87 (6)	0.77 (11)
bleurt-20	0.86 (6)	0.85 (7)	0.81 (4)
Prism	0.83 (7)	0.83 (8)	0.81 (5)
YiSi-1	0.87 (3)	0.82 (9)	0.82 (2)
C-SPECpn	0.80 (11)	0.82 (10)	0.82 (3)
COMET-MQM_2021	0.81 (9)	0.80 (11)	0.77 (9)
MTEQA	0.78 (13)	0.80 (12)	0.67 (13)
tgt-regEMT	0.78 (12)	0.80 (13)	0.82 (1)
average	0.84	0.86	0.78

Table 14: Pairwise accuracy for ranking system pairs for newstest2021 w/o HT English→German using either ref-C (main ref) or ref-A/ref-D (alternative refs), where ref-A is of substantially lower quality. ref-B was excluded because it is likely a post-edit of one of the participating systems.

8.2 Google vs. Unbabel MQM

Given that annotations were undertaken for English→Russian using a different setup and MQM scheme than those for English→German and English→Chinese we sought to provide some insight into the compatibility of the two schemes by repeating the annotation for English→German using Unbabel's scheme and annotator pool: For a subset of 5056 segments of the TED talk data for English→German from 10 MT systems, Unbabel had another expert annotator trained on MQM provide annotations using their proprietary typology. MQM was calculated for each set of annotations (using their respective scoring) and the latter were then converted to a sequence of OK/BAD tags as a means of evaluating the level of agreement between the two annotations at a token level.

The Pearson's r correlation score between the two sets of MQM annotations was found to be 0.212, significant to p<0.05. Given the levels of correlation of metrics with Google's MQM scores on the full set of English→German, this is surprisingly low. Similarly, Cohen's Kappa on the annotated tags was found to be 0.165. Not only do scores correlate poorly but agreement at the tag level is also fairly weak. Equally, Cohen's Kappa on the subset of annotations on which both sets of annotators found some error was found to be improved but still low (0.2). This indicates that even when limited to erroneous sentences, the annotators struggled to agree on where the errors were.

We note that the Google annotators left 59.5% of the sample untouched (i.e. error free), whereas the Unbabel annotator left only 46.9% untouched. It appears that the Unbabel annotator was on average more aggressive in their annotation which might partially explain low levels of agreement.

A number of the MT systems often produced the same translation of the same source text. With this in mind, and given that Google used a pool of annotators, we were able to also compare annotations within the Google set. For every source/target pair with more than two annotations we calculated and averaged the pairwise Cohen's Kappa. The mean Kappa across all of these segments was 0.21, which suggests equally low levels of agreement between Google annotators.

Despite low segment level agreement we note that the ranking of systems remains fairly consistent between annotation schemes with a few outlying exceptions. Table 15 details the rankings for our sample across annotation schemes. In particular it is encouraging to note that the human reference (albeit one of the worse ones, see Section 8.1.2) is ranked first in both cases; at a high level both schemes are making meaningful quality judgements. For the sake of completeness, we similarly examined the rankings of metrics at segment level (measuring Pearson's r correlation score and ranking the result) against both sets of MQM scores for our sample. Rankings in both cases were found to be sufficiently similar to official results reported in this paper and no metric moved more than three positions.

To rationalize these low segment-level agreement numbers, we asked an independent native language German speaker to look at a subset of annotations where we noticed the worst levels of segment-level agreement. The independent rater provided some rudimentary annotation of the most obvious errors and some qualitative analysis of the segments themselves. From this independent analysis, we were able to conclude at a high-level that the nature of TED talk text broken into segments is highly complex, context dependent and ambiguous even in the original language which resulted in equally ambiguous translation errors. This serves as a harsh reminder of the complexity of the annotation task and that inevitably even human annotation using highly granular schemes like MQM is only as reliable as the simplicity of the underlying text. The same reminder extends to human-generated references where highly specialized content will inevitably require specialized translators to ensure the most accurate translation.

System	Unbabel MQM	Google MQM
metricsystem1	88.71 (4)	-1.61 (6)
metricsystem2	87.71 (10)	-1.68 (7)
metricsystem3	86.88 (11)	-1.41 (4)
metricsystem4	87.88 (7)	-1.77 (9)
metricsystem5	87.85 (9)	-1.74 (8)
ref-A	95.49 (1)	-0.89 (1)
Facebook-AI	91.54 (3)	-1.05 (2)
Online-W	93.27 (2)	-1.12 (3)
Nemo	88.21 (6)	-2.15 (11)
VolcTrans-GLAT	88.27 (5)	-1.49 (5)
eTranslation	87.87 (8)	-1.96 (10)

Table 15: System-level MQM scores for Unbabel and Google annotation schemes

We note that whilst we do not have human direct assessment (DA) scores on TED data in order to provide a direct comparison of the two annotation schemes in this setting, we observe in the following section that MQM appears to provide a more stable basis for evaluation in general.

8.3 Comparison to WMT Scoring

The WMT evaluation campaign (Akhbardeh et al., 2021) ran a human direct assessment (DA) evaluation for the primary submissions in the news domain for all language pairs. Segment-level ratings with document context (SR+DC) on a 0-100 scale were collected either using source-based evaluation with a mix of researchers/translators (for translations out of English) or reference-based evaluation with crowd-workers (for translations into English). In general, for each MT system, only a subset of documents receive ratings, with the rated subset differing across systems. System-level DA scores are averages over the available segment-level scores. Both raw scores and per-rater z-normalized versions of the scores are provided.

Appendix C contains correlations to WMT Newstest DA scores for all metrics, at both segment and system level, for all 16 language pairs. There is significant variation in metric performance and ranking across languages, although a general pattern is that correlations are substantially higher for out-of-English pairs than into-English. Although the WMT correlations are not strictly comparable to the MQM results in previous sections, MQM scores tend to correlate somewhat better with metric scores for two of our three focus languages (English→German and Chinese→English), and somewhat worse for English→Russian.

System	MQM	WMT-raw	WMT-z
ref-C	-0.511 (1)	85.964 (5)	0.320 (3)
VolcTrans-GLAT	-1.039 (2)	87.265 (2)	0.301 (6)
Facebook-AI	-1.052 (3)	87.887 (1)	0.378 (2)
ref-A	-1.221 (4)	84.939 (9)	0.280 (8)
Nemo	-1.340 (5)	86.090 (4)	0.250 (10)
HuaweiTSC	-1.381 (6)	85.787 (6)	0.312 (4)
Online-W	-1.460 (7)	86.262 (3)	0.391 (1)
UEdin	-1.507 (8)	85.573 (8)	0.305 (5)
eTranslation	-1.695 (9)	85.706 (7)	0.281 (7)
VolcTrans-AT	-1.743 (10)	83.402 (10)	0.280 (9)

Table 16: MQM versus DA for English→German.

System	MQM	WMT-raw	WMT-z
ref-A	99.652 (1)	84.428 (1)	0.409 (1)
ref-B	98.397 (2)	83.492 (2)	0.386 (3)
Facebook-AI	92.749 (3)	81.541 (4)	0.338 (4)
Online-W	91.797 (4)	82.286 (3)	0.395 (2)
Online-A	87.866 (5)	76.177 (9)	0.227 (7)
Nemo	87.496 (6)	78.012 (7)	0.214 (8)
Online-G	87.225 (7)	78.466 (6)	0.242 (6)
Manifold	86.858 (8)	75.572 (10)	0.197 (9)
Online-B	85.663 (9)	79.962 (5)	0.294 (5)
NiuTrans	83.474 (10)	76.436 (8)	0.148 (10)
Online-Y	79.274 (11)	71.989 (11)	-0.015 (11)

Table 17: MQM versus DA for English→Russian.

System	MQM	WMT-raw	WMT-z
ref-A	-4.350 (1)	74.107 (3)	0.019 (3)
NiuTrans	-4.633 (2)	74.969 (2)	0.042 (1)
SMU	-4.844 (3)	70.559 (6)	-0.034 (7)
MiSS	-4.932 (4)	70.095 (9)	-0.029 (5)
Borderline	-4.945 (5)	70.486 (7)	-0.023 (4)
DIDI-NLP	-5.095 (6)	75.641 (1)	0.031 (2)
IIE-MT	-5.145 (7)	73.077 (4)	-0.031 (6)
Facebook-AI	-5.215 (8)	70.125 (8)	-0.037 (8)
Online-W	-5.567 (9)	72.851 (5)	-0.087 (9)

Table 18: MQM versus DA for Chinese→English.

Tables 16 to 18 compare MQM and DA scores for our focus language pairs, on all systems where both sets of scores were available. Notably, MQM scores rank human translations at or near the top more consistently than do DA scores. The only reference ranked worse than MT by MQM is ref-A for English→German, which as discussed above is a low-quality translation. In contrast, DA z-normalized scores rank all references below at least one MT system except for ref-A in English→Russian, which is ranked first, in agreement with MQM. For English→German and English→Russian, MQM correlates better with raw DA scores than with z-normalized scores; Pearson correlations are 0.508 versus 0.243 for the former and 0.911 versus 0.898 for the latter. For Chinese→English the pattern reverses, with correlations of 0.216 versus 0.729.

8.4 WMT DA as a Metric

Metric	news w/o HT	news w/ HT
BERTScore	0.902	0.097 (11)
cushLEPOR(LM)	0.898	0.023 (15)
TER	0.851	0.065 (14)
BLEU	0.850	0.090 (12)
MEE2	0.836	0.107 (9)
COMET-QE-MQM_2021-src	0.831	0.807 (1)
sentBLEU	0.824	0.114 (8)
bleurt-20	0.801	0.718 (3)
COMET-MQM_2021	0.790	0.697 (4)
Prism	0.778	-0.008 (17)
C-SPECpn	0.773	0.788 (2)
chrF	0.758	0.086 (13)
YiSi-1	0.735	0.102 (10)
regEMT	0.700	0.301 (6)
OpenKiwi-MQM-src	0.656	0.468 (5)
MTEQA	0.496	0.015 (16)
wmt-z	0.357	0.282 (7)
regEMT-src	-0.415	-0.311 (18)
YiSi-2-src	-0.609	-0.316 (19)

Table 19: System-level Pearson correlations, including WMT DA z-normalized scores as a metric, for English→German.

The correlations between MQM and WMT DA scores in the previous section motivated us to investigate how DA scores would fare in comparison to automatic metric scores when using MQM as gold scores. We computed system-level Pearson correlations using z-normalized DA scores for MT outputs only and MT outputs augmented with human references for which DA, MQM, and metric scores were all available.[16] Tables 19 to 21 compare these to the performance of primary and baseline metrics using the references from Table 7.[17] The performance of DA varies across languages: for English→German and English→Russian it ranks roughly among the

[16] ref-A for en→de, ref-B for en→ru, and ref-A for zh→en.

[17] These numbers do not match others in the paper due to the use of a reduced set of MT systems, and, for the *w/ HT* condition, a reduced set of human outputs.

Metric	news w/o HT	news w/ HT
OpenKiwi-MQM-src	0.973	0.815 (5)
C-SPECpn	0.967	0.934 (2)
Prism	0.966	-0.220 (14)
COMET-MQM_2021	0.964	0.866 (4)
BLEU	0.957	-0.025 (11)
COMET-QE-MQM_2021-src	0.953	0.946 (1)
sentBLEU	0.950	-0.011 (10)
bleurt-20	0.948	0.722 (6)
MEE2	0.937	-0.151 (12)
chrF	0.934	0.034 (9)
YiSi-1	0.932	0.079 (8)
BERTScore	0.926	-0.177 (13)
wmt-z	0.918	0.891 (3)
TER	0.841	-0.254 (15)
regEMT	0.803	0.370 (7)
regEMT-src	0.314	-0.612 (16)
YiSi-2-src	0.257	-0.652 (17)

Table 20: System-level Pearson correlations, including WMT DA z-normalized scores as a metric, for English→Russian.

Metric	news w/o HT	news w/ HT
C-SPECpn	0.797	0.882 (1)
regEMT	0.764	0.477 (5)
wmt-z	0.724	0.729 (3)
RoBLEURT	0.722	-0.237 (9)
COMET-MQM_2021	0.683	-0.034 (6)
BERTScore	0.673	-0.224 (8)
bleurt-20	0.656	-0.090 (7)
YiSi-1	0.649	-0.244 (10)
OpenKiwi-MQM-src	0.623	0.604 (4)
Prism	0.596	-0.371 (11)
COMET-QE-MQM_2021-src	0.586	0.748 (2)
chrF	0.573	-0.438 (14)
YiSi-2-src	0.519	-0.431 (13)
MEE2	0.515	-0.438 (15)
BLEU	0.507	-0.472 (16)
MTEQA	0.469	-0.424 (12)
cushLEPOR(LM)	0.460	-0.490 (18)
sentBLEU	0.441	-0.477 (17)
TER	0.316	-0.495 (19)
regEMT-src	0.004	-0.607 (20)

Table 21: System-level Pearson correlations, including WMT DA z-normalized scores as a metric, for Chinese→English.

bottom half of the automatic metrics; while for Chinese→English it ranks third. DA scores tend to perform better when judging human output, ranking 7th, 3rd, and 3rd for English→German, English→Russian, and Chinese→English, respectively.

9 Conclusion

This paper summarized the results of the WMT21 shared task on automated machine translation evaluation, the Metrics Shared Task. This year, we collected our own human ratings for selected language pairs (En→De, En→Ru, and Zh→En) from professional translators via MQM to generate a reliable ground truth across two domains. WMT direct assessment (DA) scores generally correlate poorly with MQM scores, and exhibit weaker preference for human translations compared to machine output. For En→De and Zh→En, DA ranks the human translations below many MT systems, demonstrating that expert-based evaluation is needed to generate a reliable ground truth for the Metrics Shared Task. The majority of metrics correlate better with MQM than with WMT DA, confirming previous findings that the best automatic metrics are already more reliable than crowd worker human evaluations. The performance of each metric varies depending on the underlying domain (being either TED talks or news) demonstrating that most metrics are not domain robust. Further, the challenge sets revealed that most metrics struggle to penalize translations with errors in reversing negation or sentiment polarity, and show lower correlations for Subordination, Named Entities and Terminology.

Overall, metrics perform very differently based on domain, language pair or setting (with or without human translations among candidate systems) making it hard to declare a clear winner. When counting top performances across all test conditions, three embedding-based metrics—C-SPECPN, BLEURT-20, and COMET-MQM_2021— emerge as distinctly better than the others, especially at the segment level and when rating human translations. Nevertheless, it is unclear which test scenario and correlation metric is best to yield reliable results. We would encourage the community to investigate different ways of how to evaluate automatic metrics. We are very open to apply new suggestions in the next round of the Metrics Shared Task.

Another challenge is to define the overall ground truth (i.e. the human ratings). Even though, we are convinced that expert-based ratings via MQM are more reliable, we also found that the two MQM methodologies of Unbabel and Google disagree for some systems. We would encourage the community to further work on establishing a reliable human evaluation setup. The field would benefit from a reliable human evaluation standard that could be used by everyone.

10 Ethical considerations

MQM annotations and additional reference translations in this paper are done by professional translators. They are all paid at professional rates.

Our TED talks test suite is created based on TED transcripts and translations under CC BY–NC–ND 4.0 International. Additional translations done for this shared task follow the TED Talks Usage Policy.

Organizers from the National Research Council Canada and Unbabel have submitted to this task the frozen stable versions of their metrics (YiSi and COMET) dated before they joined the organizing committee. Newer versions of COMET were developed without using any of the test set, test suite or challenge sets.

11 Acknowledgments

Results for this shared task would not be possible without tight collaboration with the organizers of the WMT News Translation Task. We are grateful to Google and Unbabel for sponsoring and overseeing the human evaluation. We would also like to thank the organizers of the WMT20 Robustness Task, especially Lucia Specia, for providing the pre-release version of the MT critical error annotations. We are grateful to Roman Grundkiewicz for his help in deciphering raw scores exported from Appraise and André Martins, Thibault Sellam, Qijun Tan and Macduff Hughes for their valuable review. Finally, we would like to thank Eleftherios Avramidis and Vivien Macketanz for providing us the German→English Challenge set (Section 6.3).

Nitika Mathur is supported by the Australian Research Council. Ricardo Rei is supported by the P2020 programs MAIA (contract 045909) and Unbabel4EU (contract 042671). Ondřej Bojar would like to acknowledge the support from the Czech Science Foundation (grant n. 19-26934X, NEUREM3).

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, and Marta R. 2021. Findings of the 2021 conference on machine translation (wmt21). In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, Aljoscha Burchardt, and Sebastian Möller. 2020. Fine-grained linguistic evaluation for state-of-the-art machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 346–356, Online. Association for Computational Linguistics.

Eleftherios Avramidis, Vivien Macketanz, Ursula Strohriegel, and Hans Uszkoreit. 2019. Linguistic evaluation of German-English machine translation using a test suite. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 445–454, Florence, Italy. 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.

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

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *arXiv preprint arXiv:2104.00054*.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *CoRR*, abs/2104.14478.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Lifeng Han, Irina Sorokina, Gleb Erofeev, and Serge Gladkoff. 2021. cushlepor: Customised hlepor metric using labse distilled knowledge model to improve agreement with human judgements. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Lifeng Han, Derek F. Wong, Lidia S. Chao, Liangye He, Yi Lu, Junwen Xing, and Xiaodong Zeng. 2013. Language-independent model for machine translation evaluation with reinforced factors. In *Machine Translation Summit XIV*, pages 215–222. International Association for Machine Translation.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. 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*, Online. Association for Computational Linguistics.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021a. Just ask! evaluating machine translation by asking and answering questions. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021b. Mteqa at wmt21 metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo. 2020. Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 895–902, Online. Association for Computational Linguistics.

Chi-kiu Lo and Samuel Larkin. 2020. Machine translation reference-less evaluation using YiSi-2 with bilingual mappings of massive multilingual language model. In *Proceedings of the Fifth Conference on Machine Translation*, pages 903–910, Online. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional Quality Metrics (MQM) : A Framework for Declaring and Describing Translation Quality Metrics. *Tradumàtica*, pages 0455–463.

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.

Vivien Macketanz, Renlong Ai, Aljoscha Burchardt, and Hans Uszkoreit. 2018a. TQ-AutoTest – an automated test suite for (machine) translation quality. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018b. Fine-grained evaluation of German-English machine translation based on a test suite. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 578–587, Belgium, Brussels. Association for Computational Linguistics.

Vivien Macketanz, Eleftherios Avramidis, Shushen Manakhimova, and Sebastian Möller. 2021. Linguistic evaluation for the 2021 state-of-the-art machine translation systems for german to english and english to german. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

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.

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.

Ananya Mukherjee, Hema Ala, Manish Shrivastava, and Dipti Misra Sharma. 2020. Mee : An automatic metric for evaluation using embeddings for machine translation. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 292–299.

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

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

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

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro G Ramos, Taisiya Glushkova, André 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*, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. 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, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. 2020. Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 921–927, Online. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Lucia Specia, Frédéric Blain, Marina Fomicheva, Erick Fonseca, Vishrav Chaudhary, Francisco Guzmán, and André F. T. Martins. 2020a. Findings of the WMT 2020 shared task on quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 743–764, Online. Association for Computational Linguistics.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020b. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Online. Association for Computational Linguistics.

Michal Stefanik, Vít Novotný, and Petr Sojka. 2021. Regressive ensemble for machine translation quality evaluation. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Kosuke Takahashi, Yoichi Ishibashi, Katsuhito Sudoh, and Satoshi Nakamura. 2021. Multilingual machine translation evaluation metrics fine-tuned on pseudo-negative examples for wmt 2021 metrics task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. 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.

Yu Wan, Dayiheng Liu, Baosong Yang, Tianchi Bi, Haibo Zhang, Boxing Chen, Weihua Luo, Derek F. Wong, and Lidia S. Chao. 2021. Robleurt submission for wmt2021 metrics task. In *Proceedings of the Sixth Conference on Machine Translation*, Online. Association for Computational Linguistics.

Evan James Williams. 1959. *Regression Analysis*, volume 14. wiley.

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

A Challenge Set Perturbation Examples

Perturbation	Description	Example
Antonym	Randomly replaces as word with it's antonym.	Orig.: *Fire in French chemical plant extinguished* New: *Fire in French chemical plant ignite*
Word omission	Randomly drops a word from a sentence	Orig.: *Fire in French chemical plant extinguished* New: *Fire in French␣ plant extinguished*
Tokenized	Applies tokenization to the sentence.	Orig.: *Spain: It is safe here.* New: *Spain␣: It is safe here␣.*
Sentence Omission	Removes a sentence from a paragraph.	Orig.: *The No.3 flood of the Yangtze River in 2020 was formed. The Ministry of Water Resources has refined and implemented countermeasures - www.chinanews.com* New: *The No.3 flood of the Yangtze River in 2020 was formed. ␣*
Punctuation	Removes punctuation from the input.	Orig.: *Spain: It is safe here.* New: *Spain␣ It is safe here␣.*
Numbers	Replaces a number by another randomly generated number.	Orig.: *Around 65 people work at the plant.* New: *Around 400 people work at the plant.*
Lowercasing	Applies lowercasing to the entire input.	Orig.: *Fire in French chemical plant extinguished* New: *fire in french chemical plant extinguished*
Word Addition	Adds a word in the middle of a sentence using `distilbert-base-uncased`. This perturbation is applied on top of lowercase perturbation.	Orig.: *fire in french chemical plant extinguished* New: *fire in french underground chemical plant extinguished*
Spelling	Adds spelling errors to the input.	Orig.: *Fire in French chemical plant extinguished* New: *Fire in French chemical pants extinguished*

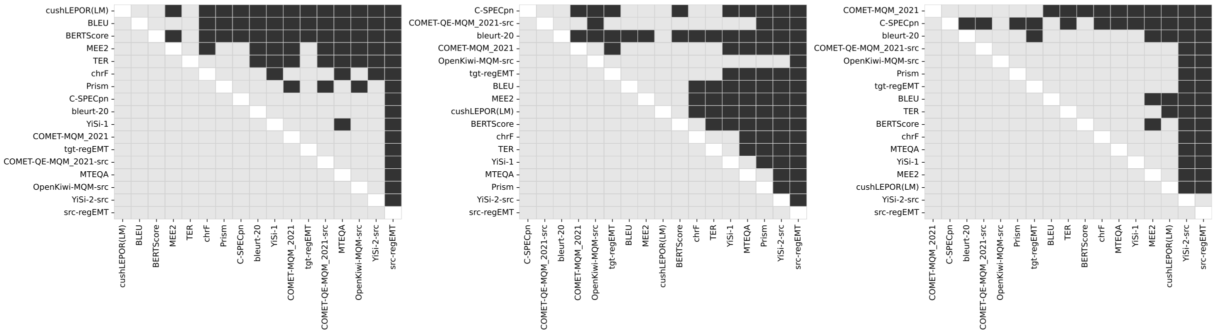
Table 22: List of all perturbations used to construct the Challenge Set described in Section 6.2. The right column provides for each perturbation an example with the original sentence and the corresponding new corrupted sentence.

B Language-Specific Results Tables

Language-specific results are given on the following pages. Each page contains results for a single language pair and granularity (system or segment). Correlation results in tables are followed by pairwise significance plots for each condition (news without human outputs, news with human outputs, TED talks) considering only primary and baseline metrics.

Metric	news w/o HT	news w/ HT	TED
cushLEPOR(LM)	**0.938** (1)	0.085 (17)	0.239 (23)
BLEU	**0.937** (2)	0.132 (13)	0.620 (13)
BERTScore	**0.930** (3)	0.074 (19)	0.506 (17)
cushLEPOR(pSQM)	0.921 (4)	0.085 (18)	0.067 (25)
MEE	0.916 (5)	0.109 (14)	0.449 (19)
MEE2	0.900 (6)	0.098 (15)	0.392 (22)
TER	0.898 (7)	0.003 (22)	0.609 (14)
hLEPOR	0.896 (8)	0.094 (16)	0.127 (24)
COMET-QE-DA_2021-src	0.847 (9)	0.807 (3)	0.527 (16)
chrF	0.846 (10)	0.017 (21)	0.471 (18)
Prism	0.841 (11)	-0.123 (26)	0.659 (11)
COMET-DA_2020	0.814 (12)	0.658 (8)	0.788 (4)
COMET-DA_2021	0.812 (13)	0.607 (9)	0.780 (5)
C-SPECpn	0.804 (14)	**0.823** (1)	**0.802** (2)
bleurt-20	0.802 (15)	**0.774** (5)	0.739 (6)
YiSi-1	0.789 (16)	-0.009 (23)	0.414 (21)
C-SPEC	0.777 (17)	0.822 (2)	0.788 (3)
COMET-MQM_2021	0.771 (18)	0.720 (7)	**0.818** (1)
bleurt-21-beta	0.771 (19)	0.758 (6)	0.695 (7)
COMETinho-DA	0.768 (20)	0.054 (20)	0.548 (15)
tgt-regEMT	0.742 (21)	0.411 (11)	0.641 (12)
COMET-QE-MQM_2021-src	0.711 (22)	**0.792** (4)	0.694 (8)
MTEQA	0.658 (23)	-0.116 (25)	0.418 (20)
tgt-regEMT-baseline	0.653 (24)	0.148 (12)	-0.078 (26)
COMETinho-MQM	0.557 (25)	-0.034 (24)	0.663 (10)
OpenKiwi-MQM-src	0.494 (26)	0.439 (10)	0.669 (9)
YiSi-2-src	0.283 (27)	-0.416 (28)	-0.419 (28)
src-regEMT-baseline	-0.173 (28)	-0.224 (27)	-0.133 (27)
src-regEMT	-0.606 (29)	-0.558 (29)	-0.699 (29)

Table 23: System-level Pearson correlations for English→German. Primary submissions are bolded, and baselines are underlined. Correlations for metrics in the top cluster (considering only primary and baseline metrics) are bolded.

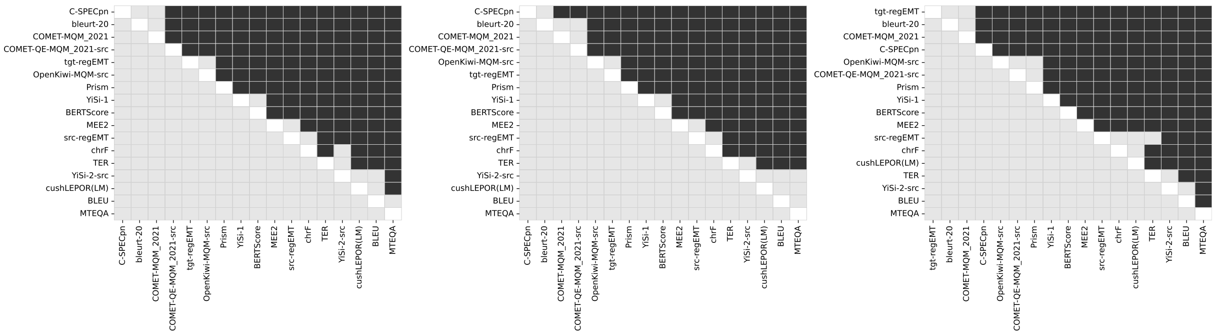


(a) news w/o HT (b) news w/ HT (c) TED w/o HT

Figure 2: System-level Pearson pairwise significance for English→German primary submissions and baselines. Dark squares mean the row metric correlates significantly better than the column metric at $\alpha = 0.05$.

Metric	news w/o HT	news w/ HT	TED
C-SPECpn	**0.267** (1)	**0.254** (2)	0.270 (5)
C-SPEC	0.266 (2)	0.256 (1)	0.285 (2)
bleurt-20	**0.264** (3)	**0.247** (3)	**0.283** (3)
COMET-MQM_2021	**0.263** (4)	0.241 (4)	**0.282** (4)
COMET-DA_2021	0.253 (5)	0.226 (8)	0.267 (6)
bleurt-21-beta	0.252 (6)	0.238 (5)	0.252 (10)
COMET-QE-MQM_2021-src	0.248 (7)	0.235 (6)	0.253 (9)
COMET-QE-DA_2021-src	0.244 (8)	0.227 (7)	0.221 (14)
COMET-DA_2020	0.239 (9)	0.212 (11)	0.259 (7)
tgt-regEMT	0.234 (10)	0.214 (10)	**0.290** (1)
OpenKiwi-MQM-src	0.232 (11)	0.219 (9)	0.255 (8)
COMETinho-MQM	0.202 (12)	0.186 (12)	0.245 (11)
COMETinho-DA	0.198 (13)	0.172 (13)	0.236 (13)
Prism	0.192 (14)	0.164 (14)	0.238 (12)
YiSi-1	0.172 (15)	0.145 (15)	0.212 (15)
BERTScore	0.169 (16)	0.143 (16)	0.189 (16)
MEE2	0.142 (17)	0.117 (17)	0.173 (17)
src-regEMT	0.128 (18)	0.106 (18)	0.149 (19)
MEE	0.126 (19)	0.105 (19)	0.142 (22)
chrF	0.114 (20)	0.090 (20)	0.146 (20)
TER	0.098 (21)	0.078 (22)	0.131 (23)
YiSi-2-src	0.098 (22)	0.071 (23)	0.119 (25)
cushLEPOR(LM)	0.090 (23)	0.068 (24)	0.144 (21)
tgt-regEMT-baseline	0.084 (24)	0.080 (21)	0.161 (18)
sentBLEU	0.083 (25)	0.064 (26)	0.113 (27)
cushLEPOR(pSQM)	0.078 (26)	0.057 (28)	0.127 (24)
MTEQA	0.071 (27)	0.060 (27)	0.082 (29)
hLEPOR	0.071 (28)	0.050 (29)	0.117 (26)
src-regEMT-baseline	0.067 (29)	0.067 (25)	0.112 (28)

Table 24: Segment-level Kendall correlations for English→German. Primary submissions are bolded, and baselines are underlined. Correlations for metrics in the top cluster (considering only primary and baseline metrics) are bolded.



(a) news w/o HT

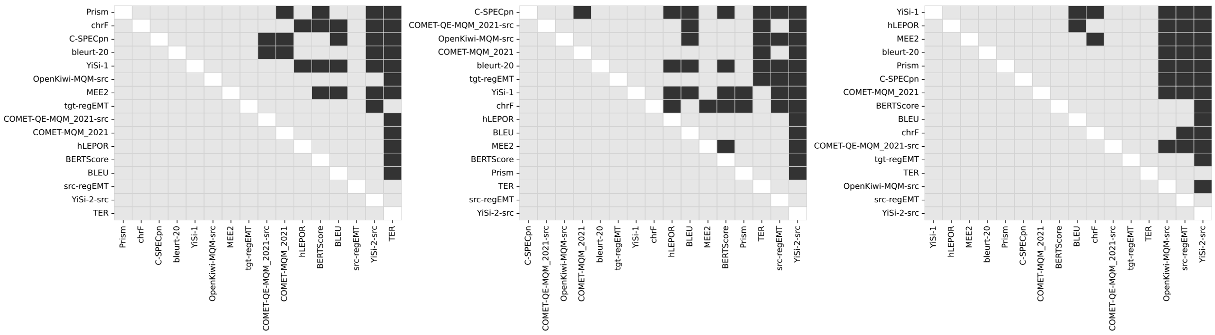
(b) news w/ HT

(c) TED w/o HT

Figure 3: Segment-level Kendall significance for English→German primary submissions and baselines. Dark squares mean the row metric correlates significantly better than the column metric at $\alpha = 0.05$.

Metric	news w/o HT	news w/ HT	TED
Prism	**0.799** (1)	-0.136 (21)	**0.867** (8)
chrF	**0.783** (2)	0.123 (14)	0.825 (16)
C-SPECpn	**0.782** (3)	**0.824** (1)	**0.855** (12)
bleurt-20	**0.768** (4)	0.653 (8)	**0.868** (7)
C-SPEC	0.763 (5)	0.817 (2)	0.858 (10)
YiSi-1	**0.761** (6)	0.138 (13)	**0.905** (1)
MEE	0.759 (7)	0.051 (15)	0.881 (5)
OpenKiwi-MQM-src	**0.755** (8)	**0.729** (4)	0.691 (21)
MEE2	**0.750** (9)	-0.069 (18)	**0.882** (4)
bleurt-21-beta	0.743 (10)	0.692 (5)	0.856 (11)
tgt-regEMT	**0.740** (11)	0.390 (11)	0.758 (19)
COMET-QE-MQM_2021-src	0.688 (12)	**0.784** (3)	0.817 (17)
COMET-DA_2020	0.676 (13)	0.556 (10)	0.859 (9)
COMET-MQM_2021	0.659 (14)	0.685 (6)	**0.841** (13)
COMET-DA_2021	0.655 (15)	0.645 (9)	0.871 (6)
hLEPOR	0.648 (16)	-0.038 (16)	**0.894** (2)
COMET-QE-DA_2021-src	0.632 (17)	0.681 (7)	0.884 (3)
BERTScore	0.629 (18)	-0.123 (20)	**0.831** (14)
COMETinho-DA	0.578 (19)	0.239 (12)	0.758 (18)
BLEU	0.507 (20)	-0.043 (17)	0.828 (15)
src-regEMT	0.301 (21)	-0.436 (24)	0.115 (24)
tgt-regEMT-baseline	0.186 (22)	-0.413 (23)	0.121 (23)
COMETinho-MQM	0.089 (23)	-0.083 (19)	0.432 (22)
YiSi-2-src	0.046 (24)	-0.585 (26)	0.085 (25)
TER	-0.041 (25)	-0.289 (22)	0.697 (20)
src-regEMT-baseline	-0.585 (26)	-0.583 (25)	-0.228 (26)

Table 25: System-level Pearson correlations for English→Russian. Primary submissions are bolded, and baselines are underlined. Correlations for metrics in the top cluster are bolded.

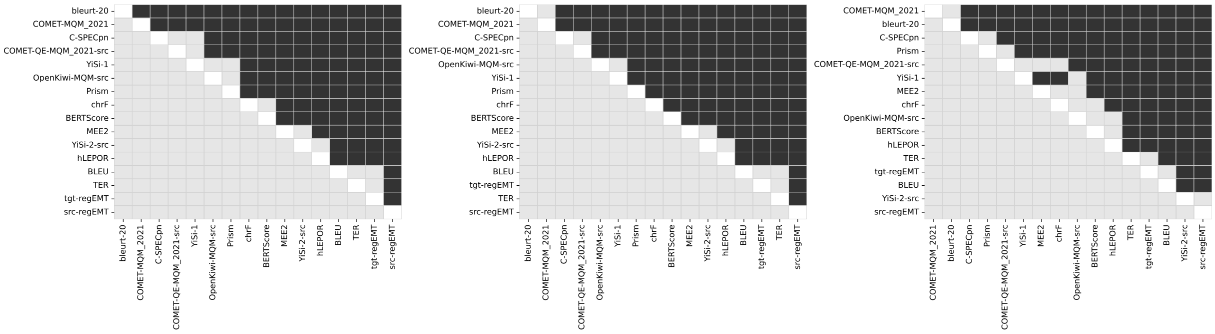


(a) news w/o HT (b) news w/ HT (c) TED w/o HT

Figure 4: System-level Pearson pairwise significance for English→Russian primary submissions and baselines. Dark squares mean the row metric correlates significantly better than the column metric at $\alpha = 0.05$.

Metric	news w/o HT	news w/ HT	TED
COMET-DA_2021	0.307 (1)	0.296 (1)	0.274 (1)
bleurt-20	**0.286** (2)	**0.276** (3)	**0.255** (5)
COMET-QE-DA_2021-src	0.284 (3)	0.278 (2)	0.245 (6)
bleurt-21-beta	0.278 (4)	0.271 (4)	0.269 (2)
COMET-DA_2020	0.278 (5)	0.265 (6)	0.242 (7)
COMET-MQM_2021	0.276 (6)	**0.268** (5)	**0.258** (4)
C-SPEC	0.259 (7)	0.259 (7)	0.263 (3)
C-SPECpn	0.248 (8)	0.248 (8)	0.233 (8)
COMETinho-DA	0.248 (9)	0.233 (10)	0.218 (10)
COMET-QE-MQM_2021-src	0.242 (10)	0.239 (9)	0.204 (12)
YiSi-1	0.233 (11)	0.216 (12)	0.204 (11)
OpenKiwi-MQM-src	0.225 (12)	0.222 (11)	0.187 (15)
<u>Prism</u>	0.224 (13)	0.205 (13)	0.219 (9)
COMETinho-MQM	0.197 (14)	0.188 (14)	0.182 (17)
<u>chrF</u>	0.193 (15)	0.178 (15)	0.189 (14)
<u>BERTScore</u>	0.185 (16)	0.168 (16)	0.185 (16)
MEE2	0.169 (17)	0.153 (17)	0.193 (13)
YiSi-2-src	0.163 (18)	0.140 (18)	0.084 (23)
MEE	0.150 (19)	0.135 (20)	0.176 (19)
hLEPOR	0.150 (20)	0.135 (19)	0.178 (18)
<u>sentBLEU</u>	0.120 (21)	0.106 (21)	0.112 (22)
<u>TER</u>	0.117 (22)	0.104 (23)	0.142 (20)
tgt-regEMT	0.110 (23)	0.105 (22)	0.129 (21)
src-regEMT	0.085 (24)	0.070 (24)	0.070 (24)
tgt-regEMT-baseline	0.053 (25)	0.050 (25)	0.053 (25)
src-regEMT-baseline	-0.045 (26)	-0.043 (26)	0.018 (26)

Table 26: Segment-level Kendall correlations for English→Russian. Primary submissions are bolded, and baselines are underlined. Correlations for metrics in the top cluster (considering only primary and baseline metrics) are bolded.



(a) news w/o HT

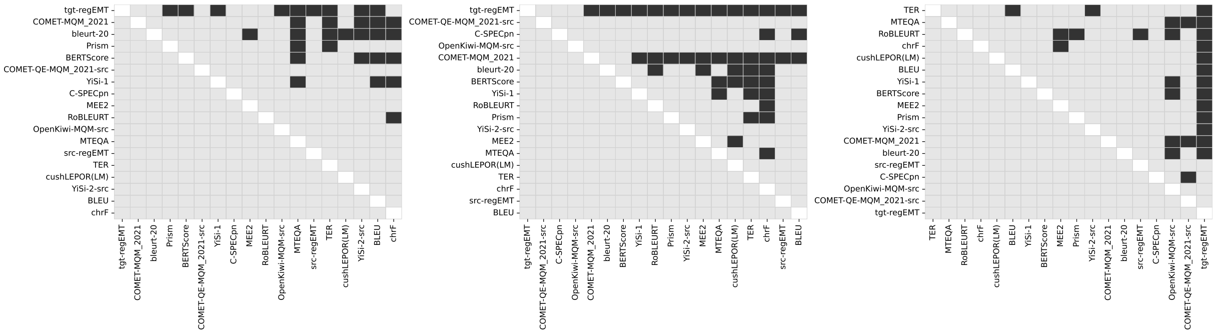
(b) news w/ HT

(c) TED w/o HT

Figure 5: Segment-level Kendall significance for English→Russian primary submissions and baselines. Dark squares mean the row metric correlates significantly better than the column metric at $\alpha = 0.05$.

Metric	news w/o HT	news w/ HT	TED
tgt-regEMT	**0.834** (1)	**0.727** (1)	-0.404 (30)
COMET-MQM_2021	**0.628** (2)	0.336 (7)	0.266 (18)
bleurt-20	**0.563** (3)	0.294 (9)	0.239 (20)
Prism	0.558 (4)	0.031 (17)	0.272 (15)
BERTScore	0.542 (5)	0.095 (14)	0.306 (12)
bleurt-21-beta	0.537 (6)	0.265 (10)	0.235 (21)
COMETinho-MQM	0.530 (7)	0.129 (13)	0.395 (5)
COMET-QE-MQM_2021-src	0.529 (8)	**0.619** (2)	-0.209 (29)
C-SPEC	0.526 (9)	0.619 (3)	-0.064 (26)
COMET-DA_2021	0.516 (10)	0.186 (12)	0.306 (11)
YiSi-1	0.515 (11)	0.077 (15)	0.310 (10)
COMET-DA_2020	0.511 (12)	0.221 (11)	0.251 (19)
hLEPOR	0.498 (13)	-0.061 (24)	0.372 (6)
C-SPECpn	0.492 (14)	**0.594** (4)	-0.053 (25)
MEE2	0.453 (15)	-0.011 (19)	0.289 (14)
COMET-QE-DA_2021-src	0.453 (16)	0.535 (5)	0.057 (24)
RoBLEURT	0.451 (17)	0.065 (16)	**0.400** (3)
OpenKiwi-MQM-src	0.445 (18)	**0.489** (6)	-0.077 (27)
MTEQA	0.423 (19)	-0.050 (21)	**0.403** (2)
src-regEMT	0.419 (20)	-0.149 (29)	0.077 (23)
TER	0.416 (21)	-0.085 (26)	**0.421** (1)
cushLEPOR(LM)	0.412 (22)	-0.052 (22)	**0.327** (8)
YiSi-2-src	0.411 (23)	0.013 (18)	0.270 (16)
COMETinho-DA	0.340 (24)	-0.019 (20)	0.397 (4)
MEE	0.324 (25)	-0.125 (27)	0.301 (13)
src-regEMT-baseline	0.310 (26)	0.300 (8)	-0.105 (28)
BLEU	0.310 (27)	-0.152 (30)	0.324 (9)
chrF	0.302 (28)	-0.143 (28)	**0.363** (7)
cushLEPOR(pSQM)	0.237 (29)	-0.058 (23)	0.267 (17)
tgt-regEMT-baseline	0.089 (30)	-0.075 (25)	0.201 (22)

Table 27: System-level Pearson correlations for Chinese→English. Primary submissions are bolded, and baselines are underlined. Correlations for metrics in the top cluster are bolded.



(a) news w/o HT

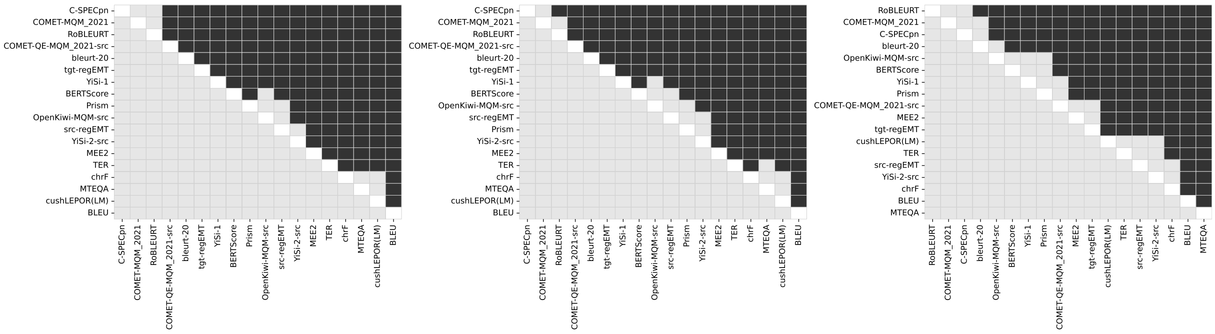
(b) news w/ HT

(c) TED w/o HT

Figure 6: System-level Pearson pairwise significance for Chinese→English primary submissions and baselines. Dark squares mean the row metric correlates significantly better than the column metric at $\alpha = 0.05$.

Metric	news w/o HT	news w/ HT	TED
C-SPECpn	**0.402** (1)	**0.390** (1)	**0.233** (7)
C-SPEC	0.401 (2)	0.388 (2)	0.241 (2)
COMET-MQM_2021	**0.395** (3)	**0.384** (3)	**0.233** (5)
RoBLEURT	**0.394** (4)	0.380 (4)	**0.238** (4)
COMET-DA_2021	0.371 (5)	0.357 (7)	0.219 (11)
COMETinho-MQM	0.370 (6)	0.362 (5)	0.239 (3)
COMET-QE-MQM_2021-src	0.367 (7)	0.358 (6)	0.178 (17)
COMET-DA_2020	0.360 (8)	0.347 (8)	0.220 (10)
bleurt-21-beta	0.357 (9)	0.344 (9)	0.233 (6)
bleurt-20	0.354 (10)	0.341 (10)	0.224 (9)
COMETinho-DA	0.339 (11)	0.327 (11)	0.199 (14)
tgt-regEMT	0.328 (12)	0.318 (12)	0.173 (18)
COMET-QE-DA_2021-src	0.305 (13)	0.294 (13)	0.122 (28)
YiSi-1	0.302 (14)	0.289 (14)	0.195 (15)
BERTScore	0.296 (15)	0.281 (15)	0.199 (13)
Prism	0.285 (16)	0.270 (19)	0.194 (16)
OpenKiwi-MQM-src	0.283 (17)	0.277 (16)	0.213 (12)
src-regEMT	0.280 (18)	0.274 (17)	0.135 (23)
tgt-regEMT-baseline	0.278 (19)	0.272 (18)	0.248 (1)
YiSi-2-src	0.270 (20)	0.263 (20)	0.125 (26)
src-regEMT-baseline	0.255 (21)	0.251 (21)	0.231 (8)
MEE2	0.247 (22)	0.233 (22)	0.173 (19)
TER	0.210 (23)	0.198 (23)	0.136 (22)
hLEPOR	0.205 (24)	0.193 (24)	0.129 (25)
chrF	0.201 (25)	0.188 (25)	0.124 (27)
MEE	0.196 (26)	0.186 (27)	0.131 (24)
MTEQA	0.194 (27)	0.187 (26)	0.028 (30)
cushLEPOR(LM)	0.193 (28)	0.182 (28)	0.138 (21)
sentBLEU	0.176 (29)	0.165 (29)	0.092 (29)
cushLEPOR(pSQM)	0.167 (30)	0.158 (30)	0.143 (20)

Table 28: Segment-level Kendall correlations for Chinese→English. Primary submissions are bolded, and baselines are underlined. Correlations for metrics in the top cluster are bolded.



(a) news w/o HT

(b) news w/ HT

(c) TED w/o HT

Figure 7: Segment-level Kendall significance for Chinese→English primary submissions and baselines. Dark squares mean the row metric correlates significantly better than the column metric at $\alpha = 0.05$.

C WMT Direct Assessment Results

Correlations with WMT Direct Assessment scores for the news and FLORES test sets are given in the following tables, with results for news to-English language pairs followed by to-non-English pairs, followed by FLORES. Since most language pairs contained only a single reference, we used reference A for all pairs, and report results only for scoring MT output (omitting additional scored references for language pairs where these were available). System-level correlations use Pearson over z-normalized rater scores. Segment-level correlations use the traditional Kendall-like formula over raw rater scores, discarding segment pairs whose scores differ by less than 25.[18]

metric	correlation
regEMT-src	0.778
COMETinho-MQM	0.652
Prism	0.651
RoBLEURT	0.648
OpenKiwi-MQM-src	0.641
COMET-MQM_2021	0.638
COMET-DA_2020	0.632
BERTScore	0.629
bleurt-21-beta	0.628
COMET-DA_2021	0.626
COMET-QE-MQM_2021-src	0.625
C-SPEC	0.623
bleurt-20	0.620
regEMT	0.609
YiSi-1	0.607
COMET-QE-DA_2021-src	0.606
C-SPECpn	0.590
COMETinho-DA	0.588
MTEQA	0.586
chrF	0.562
sentBLEU	0.550
TER	0.509
hLEPOR	0.496
YiSi-2-src	0.248
regEMT-baseline	-0.195
regEMT-baseline-src	-0.335

metric	correlation
RoBLEURT	0.044
COMET-MQM_2021	0.037
COMETinho-MQM	0.034
COMET-QE-MQM_2021-src	0.033
COMET-DA_2021	0.032
COMET-DA_2020	0.032
regEMT	0.027
COMET-QE-DA_2021-src	0.026
OpenKiwi-MQM-src	0.018
YiSi-2-src	0.017
COMETinho-DA	0.015
C-SPECpn	0.008
regEMT-src	0.003
Prism	-0.002
C-SPEC	-0.012
bleurt-20	-0.017
BERTScore	-0.019
bleurt-21-beta	-0.026
YiSi-1	-0.039
chrF	-0.053
sentBLEU	-0.088
hLEPOR	-0.098
regEMT-baseline	-0.118
regEMT-baseline-src	-0.135
TER	-0.226
MTEQA	-0.237

Table 29: Correlations for Czech→English: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

[18] Note that we used average sentence-level BLEU rather than corpus BLEU for system-level results, in contrast to our main results.

metric	correlation
regEMT-baseline	0.520
hLEPOR	0.493
YiSi-1	0.395
MTEQA	0.394
regEMT	0.362
COMET-DA_2020	0.361
chrF	0.357
YiSi-2-src	0.354
COMET-DA_2021	0.354
RoBLEURT	0.353
Prism	0.349
COMET-MQM_2021	0.346
bleurt-20	0.340
BERTScore	0.336
COMETinho-DA	0.333
bleurt-21-beta	0.325
COMET-QE-DA_2021-src	0.320
COMET-QE-MQM_2021-src	0.293
sentBLEU	0.231
OpenKiwi-MQM-src	0.215
COMETinho-MQM	0.163
C-SPECpn	0.122
C-SPEC	0.090
TER	0.070
regEMT-src	0.064
regEMT-baseline-src	-0.499

metric	correlation
RoBLEURT	0.011
COMET-QE-MQM_2021-src	0.004
COMETinho-DA	0.001
COMET-DA_2020	-0.002
YiSi-2-src	-0.003
COMET-QE-DA_2021-src	-0.003
COMET-MQM_2021	-0.003
COMETinho-MQM	-0.005
COMET-DA_2021	-0.006
OpenKiwi-MQM-src	-0.020
regEMT	-0.025
regEMT-src	-0.034
Prism	-0.037
C-SPECpn	-0.091
C-SPEC	-0.093
BERTScore	-0.098
bleurt-20	-0.146
YiSi-1	-0.151
bleurt-21-beta	-0.153
chrF	-0.162
hLEPOR	-0.209
sentBLEU	-0.215
regEMT-baseline	-0.231
regEMT-baseline-src	-0.234
TER	-0.340
MTEQA	-0.413

Table 30: Correlations for German→English: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
bleurt-20	0.955
COMET-DA_2020	0.949
Prism	0.948
bleurt-21-beta	0.947
BERTScore	0.947
YiSi-1	0.944
RoBLEURT	0.944
regEMT	0.940
COMET-DA_2021	0.939
sentBLEU	0.936
chrF	0.924
COMETinho-DA	0.923
MTEQA	0.909
COMET-MQM_2021	0.902
COMET-QE-DA_2021-src	0.898
COMETinho-MQM	0.880
TER	0.823
C-SPEC	0.810
OpenKiwi-MQM-src	0.806
YiSi-2-src	0.795
COMET-QE-MQM_2021-src	0.782
C-SPECpn	0.720
regEMT-baseline	0.525
regEMT-src	0.363
regEMT-baseline-src	0.014

metric	correlation
COMET-MQM_2021	0.076
RoBLEURT	0.075
COMET-DA_2021	0.072
C-SPEC	0.070
Prism	0.070
C-SPECpn	0.066
COMET-QE-DA_2021-src	0.064
COMET-DA_2020	0.062
BERTScore	0.062
COMETinho-DA	0.056
OpenKiwi-MQM-src	0.051
YiSi-1	0.049
COMET-QE-MQM_2021-src	0.047
bleurt-20	0.046
YiSi-2-src	0.046
regEMT	0.043
bleurt-21-beta	0.039
COMETinho-MQM	0.036
chrF	0.021
regEMT-src	0.009
sentBLEU	-0.010
regEMT-baseline	-0.067
regEMT-baseline-src	-0.067
MTEQA	-0.067
TER	-0.125

Table 31: Correlations for Hausa→English: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
RoBLEURT	0.891
bleurt-21-beta	0.889
bleurt-20	0.888
COMET-QE-MQM_2021-src	0.887
OpenKiwi-MQM-src	0.879
COMET-MQM_2021	0.872
TER	0.869
YiSi-1	0.868
BERTScore	0.867
COMET-DA_2021	0.866
sentBLEU	0.858
COMET-QE-DA_2021-src	0.857
regEMT	0.856
chrF	0.854
COMET-DA_2020	0.849
Prism	0.846
MTEQA	0.831
COMETinho-DA	0.818
COMETinho-MQM	0.800
regEMT-src	0.665
regEMT-baseline-src	0.632
YiSi-2-src	0.628
C-SPECpn	0.622
regEMT-baseline	0.445
C-SPEC	-0.104

metric	correlation
COMET-MQM_2021	0.069
Prism	0.063
RoBLEURT	0.063
COMET-DA_2021	0.061
COMET-QE-MQM_2021-src	0.061
COMET-DA_2020	0.057
C-SPEC	0.057
COMETinho-DA	0.055
COMET-QE-DA_2021-src	0.051
COMETinho-MQM	0.048
regEMT	0.041
C-SPECpn	0.041
BERTScore	0.038
YiSi-2-src	0.035
OpenKiwi-MQM-src	0.031
bleurt-20	0.030
bleurt-21-beta	0.028
regEMT-src	0.027
YiSi-1	0.023
chrF	0.018
sentBLEU	-0.018
regEMT-baseline	-0.063
regEMT-baseline-src	-0.083
TER	-0.126
MTEQA	-0.157

Table 32: Correlations for Icelandic→English: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
COMET-DA_2020	0.846
COMETinho-DA	0.839
COMET-DA_2021	0.832
chrF	0.831
Prism	0.827
COMETinho-MQM	0.824
YiSi-1	0.821
RoBLEURT	0.820
BERTScore	0.819
bleurt-20	0.806
bleurt-21-beta	0.803
sentBLEU	0.787
MTEQA	0.784
COMET-MQM_2021	0.766
COMET-QE-DA_2021-src	0.759
regEMT	0.739
regEMT-baseline	0.716
YiSi-2-src	0.696
TER	0.693
OpenKiwi-MQM-src	0.584
COMET-QE-MQM_2021-src	0.567
C-SPEC	0.365
regEMT-src	0.071
C-SPECpn	-0.074
regEMT-baseline-src	-0.710

metric	correlation
RoBLEURT	0.045
Prism	0.035
COMET-DA_2020	0.033
COMET-MQM_2021	0.032
COMET-DA_2021	0.031
C-SPEC	0.030
BERTScore	0.028
COMETinho-DA	0.025
COMET-QE-DA_2021-src	0.025
C-SPECpn	0.024
YiSi-1	0.022
OpenKiwi-MQM-src	0.021
COMET-QE-MQM_2021-src	0.012
regEMT	0.009
YiSi-2-src	0.009
bleurt-20	0.007
chrF	0.005
COMETinho-MQM	0.002
bleurt-21-beta	0.002
regEMT-src	-0.004
sentBLEU	-0.023
regEMT-baseline	-0.054
regEMT-baseline-src	-0.070
MTEQA	-0.082
TER	-0.129

Table 33: Correlations for Japanese→English: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
COMET-QE-DA_2021-src	0.764
OpenKiwi-MQM-src	0.759
regEMT	0.748
COMET-QE-MQM_2021-src	0.742
bleurt-21-beta	0.732
COMET-MQM_2021	0.728
bleurt-20	0.728
COMET-DA_2020	0.726
COMET-DA_2021	0.711
MTEQA	0.705
RoBLEURT	0.687
regEMT-baseline	0.676
BERTScore	0.668
COMETinho-MQM	0.658
Prism	0.657
YiSi-1	0.652
COMETinho-DA	0.627
chrF	0.593
hLEPOR	0.527
sentBLEU	0.512
TER	0.481
C-SPEC	0.456
C-SPECpn	0.394
YiSi-2-src	0.335
regEMT-src	0.092
regEMT-baseline-src	-0.535

metric	correlation
OpenKiwi-MQM-src	0.024
COMET-QE-MQM_2021-src	0.018
regEMT	0.017
COMET-QE-DA_2021-src	0.007
COMET-MQM_2021	0.005
COMET-DA_2021	-0.006
RoBLEURT	-0.006
C-SPECpn	-0.017
YiSi-2-src	-0.017
regEMT-src	-0.017
COMETinho-DA	-0.021
COMET-DA_2020	-0.022
COMETinho-MQM	-0.023
C-SPEC	-0.029
Prism	-0.033
BERTScore	-0.081
bleurt-20	-0.105
bleurt-21-beta	-0.109
chrF	-0.126
YiSi-1	-0.127
regEMT-baseline-src	-0.139
sentBLEU	-0.144
hLEPOR	-0.144
regEMT-baseline	-0.167
TER	-0.263
MTEQA	-0.314

Table 34: Correlations for Russian→English: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
COMET-MQM_2021	0.762
COMET-DA_2021	0.756
bleurt-21-beta	0.754
bleurt-20	0.749
COMET-DA_2020	0.748
COMET-QE-DA_2021-src	0.746
YiSi-1	0.735
COMET-QE-MQM_2021-src	0.731
BERTScore	0.727
MEE	0.726
Prism	0.726
chrF	0.723
RoBLEURT	0.720
regEMT	0.712
sentBLEU	0.709
MEE2	0.707
OpenKiwi-MQM-src	0.706
regEMT-baseline	0.699
COMETinho-DA	0.693
cushLEPOR(pSQM)	0.679
cushLEPOR(LM)	0.678
MTEQA	0.661
BLEU	0.653
COMETinho-MQM	0.568
hLEPOR	0.550
YiSi-2-src	0.542
TER	0.527
regEMT-src	0.378
C-SPEC	0.218
C-SPECpn	0.214
regEMT-baseline-src	-0.669

metric	correlation
OpenKiwi-MQM-src	0.021
COMET-MQM_2021	0.020
COMET-QE-MQM_2021-src	0.020
RoBLEURT	0.019
COMET-DA_2021	0.018
COMETinho-DA	0.018
COMET-QE-DA_2021-src	0.017
COMET-DA_2020	0.016
YiSi-2-src	0.008
COMETinho-MQM	0.008
Prism	0.007
regEMT-src	-0.005
C-SPECpn	-0.005
regEMT	-0.006
C-SPEC	-0.007
BERTScore	-0.013
bleurt-21-beta	-0.022
YiSi-1	-0.026
bleurt-20	-0.028
MEE2	-0.035
chrF	-0.035
hLEPOR	-0.050
cushLEPOR(LM)	-0.050
cushLEPOR(pSQM)	-0.056
sentBLEU	-0.057
MEE	-0.063
regEMT-baseline	-0.089
regEMT-baseline-src	-0.090
MTEQA	-0.118
TER	-0.165

Table 35: Correlations for Chinese→English: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
YiSi-1	0.781
COMET-DA_2021	0.774
COMET-DA_2020	0.743
bleurt-20	0.734
bleurt-21-beta	0.721
COMET-MQM_2021	0.693
COMET-QE-DA_2021-src	0.658
COMETinho-DA	0.620
Prism	0.584
regEMT	0.534
regEMT-baseline	0.526
COMET-QE-MQM_2021-src	0.516
OpenKiwi-MQM-src	0.460
YiSi-2-src	0.414
chrF	0.413
BERTScore	0.378
TER	0.363
sentBLEU	0.363
C-SPECpn	0.329
C-SPEC	0.320
COMETinho-MQM	0.298
regEMT-src	0.101
regEMT-baseline-src	-0.433

metric	correlation
COMET-MQM_2021	0.223
COMET-DA_2021	0.220
Prism	0.208
COMET-QE-DA_2021-src	0.203
bleurt-20	0.202
COMET-DA_2020	0.202
bleurt-21-beta	0.193
C-SPEC	0.189
YiSi-1	0.173
COMET-QE-MQM_2021-src	0.161
COMETinho-DA	0.156
C-SPECpn	0.143
OpenKiwi-MQM-src	0.123
COMETinho-MQM	0.118
BERTScore	0.116
chrF	0.110
regEMT	0.104
YiSi-2-src	0.104
sentBLEU	0.055
regEMT-src	0.041
regEMT-baseline	0.031
TER	-0.063
regEMT-baseline-src	-0.201

Table 36: Correlations for German→French: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
bleurt-21-beta	0.991
bleurt-20	0.989
YiSi-1	0.985
COMET-MQM_2021	0.979
COMET-DA_2021	0.979
sentBLEU	0.976
chrF	0.975
COMET-QE-MQM_2021-src	0.974
Prism	0.971
COMET-DA_2020	0.971
regEMT	0.969
TER	0.967
COMET-QE-DA_2021-src	0.966
BERTScore	0.965
hLEPOR	0.956
COMETinho-DA	0.941
MTEQA	0.905
COMETinho-MQM	0.895
OpenKiwi-MQM-src	0.873
regEMT-baseline	0.866
regEMT-src	0.595
C-SPEC	0.123
C-SPECpn	0.072
YiSi-2-src	-0.007
regEMT-baseline-src	-0.920

metric	correlation
COMET-DA_2021	0.774
bleurt-20	0.764
COMET-MQM_2021	0.757
C-SPEC	0.753
bleurt-21-beta	0.752
COMET-DA_2020	0.737
COMET-QE-DA_2021-src	0.724
C-SPECpn	0.723
COMET-QE-MQM_2021-src	0.714
Prism	0.712
YiSi-1	0.686
OpenKiwi-MQM-src	0.652
regEMT	0.641
COMETinho-DA	0.573
BERTScore	0.571
chrF	0.531
COMETinho-MQM	0.492
hLEPOR	0.441
sentBLEU	0.383
YiSi-2-src	0.240
MTEQA	0.212
TER	0.208
regEMT-src	0.160
regEMT-baseline	0.126
regEMT-baseline-src	-0.349

Table 37: Correlations for English→Czech: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
bleurt-20	0.885
bleurt-21-beta	0.882
COMET-MQM_2021	0.880
COMET-QE-MQM_2021-src	0.877
COMET-DA_2021	0.875
OpenKiwi-MQM-src	0.870
COMET-QE-DA_2021-src	0.866
COMET-DA_2020	0.864
regEMT	0.855
COMETinho-DA	0.854
Prism	0.853
YiSi-1	0.847
COMETinho-MQM	0.835
chrF	0.820
MTEQA	0.797
TER	0.794
BERTScore	0.794
MEE2	0.793
sentBLEU	0.769
MEE	0.761
BLEU	0.738
cushLEPOR(pSQM)	0.699
cushLEPOR(LM)	0.691
hLEPOR	0.671
regEMT-src	0.481
C-SPECpn	0.408
C-SPEC	0.258
YiSi-2-src	0.025
regEMT-baseline	-0.272
regEMT-baseline-src	-0.727

metric	correlation
COMET-DA_2021	0.255
COMET-DA_2020	0.255
COMET-MQM_2021	0.247
COMET-QE-DA_2021-src	0.237
COMET-QE-MQM_2021-src	0.230
regEMT	0.220
Prism	0.208
OpenKiwi-MQM-src	0.205
bleurt-21-beta	0.202
C-SPEC	0.200
bleurt-20	0.200
C-SPECpn	0.199
YiSi-1	0.162
COMETinho-DA	0.162
COMETinho-MQM	0.157
BERTScore	0.146
MEE2	0.102
chrF	0.098
YiSi-2-src	0.075
regEMT-src	0.067
cushLEPOR(LM)	0.033
hLEPOR	0.026
cushLEPOR(pSQM)	0.025
MEE	0.019
sentBLEU	0.014
MTEQA	-0.122
TER	-0.123
regEMT-baseline	-0.136
regEMT-baseline-src	-0.180

Table 38: Correlations for English→German: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
bleurt-20	0.915
bleurt-21-beta	0.907
regEMT	0.901
YiSi-1	0.892
COMET-DA_2020	0.871
COMET-DA_2021	0.863
BERTScore	0.838
COMET-MQM_2021	0.811
OpenKiwi-MQM-src	0.791
sentBLEU	0.789
COMET-QE-DA_2021-src	0.786
chrF	0.768
COMET-QE-MQM_2021-src	0.746
COMETinho-DA	0.708
COMETinho-MQM	0.463
regEMT-baseline	0.376
YiSi-2-src	0.362
TER	0.288
C-SPEC	0.174
C-SPECpn	0.077
regEMT-src	-0.266
regEMT-baseline-src	-0.357

metric	correlation
COMET-DA_2021	0.237
COMET-DA_2020	0.234
COMET-MQM_2021	0.214
C-SPEC	0.210
COMET-QE-DA_2021-src	0.198
bleurt-20	0.186
C-SPECpn	0.186
chrF	0.186
bleurt-21-beta	0.183
YiSi-1	0.180
COMET-QE-MQM_2021-src	0.176
BERTScore	0.167
OpenKiwi-MQM-src	0.157
COMETinho-DA	0.131
regEMT	0.130
sentBLEU	0.124
YiSi-2-src	0.102
COMETinho-MQM	0.088
regEMT-baseline	0.049
regEMT-src	0.016
TER	-0.025
regEMT-baseline-src	-0.112

Table 39: Correlations for English→Hausa: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
regEMT	0.989
bleurt-20	0.975
sentBLEU	0.962
bleurt-21-beta	0.962
chrF	0.961
COMET-MQM_2021	0.960
COMET-DA_2021	0.959
COMET-QE-MQM_2021-src	0.959
YiSi-1	0.957
COMET-DA_2020	0.952
BERTScore	0.950
COMET-QE-DA_2021-src	0.945
COMETinho-DA	0.931
TER	0.928
COMETinho-MQM	0.908
OpenKiwi-MQM-src	0.873
C-SPECpn	0.750
C-SPEC	0.736
regEMT-baseline	0.478
YiSi-2-src	0.348
regEMT-src	0.125
regEMT-baseline-src	-0.922

metric	correlation
COMET-MQM_2021	0.489
COMET-DA_2021	0.487
COMET-DA_2020	0.474
C-SPEC	0.472
bleurt-20	0.469
C-SPECpn	0.460
COMET-QE-DA_2021-src	0.454
COMET-QE-MQM_2021-src	0.453
bleurt-21-beta	0.444
YiSi-1	0.410
OpenKiwi-MQM-src	0.404
COMETinho-DA	0.384
chrF	0.373
BERTScore	0.355
COMETinho-MQM	0.330
regEMT	0.312
sentBLEU	0.279
TER	0.121
YiSi-2-src	0.105
regEMT-src	0.012
regEMT-baseline	0.002
regEMT-baseline-src	-0.199

Table 40: Correlations for English→Icelandic: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
bleurt-21-beta	0.991
COMET-DA_2020	0.988
bleurt-20	0.985
COMET-DA_2021	0.984
COMET-QE-DA_2021-src	0.978
YiSi-1	0.974
COMETinho-DA	0.972
regEMT	0.972
Prism	0.971
COMET-MQM_2021	0.970
chrF	0.966
COMETinho-MQM	0.955
COMET-QE-MQM_2021-src	0.947
BERTScore	0.939
OpenKiwi-MQM-src	0.929
C-SPECpn	0.678
regEMT-src	0.502
YiSi-2-src	0.470
regEMT-baseline	0.423
C-SPEC	0.325
TER	-0.025
regEMT-baseline-src	-0.216
sentBLEU	-0.629

metric	correlation
COMET-DA_2021	0.531
COMET-DA_2020	0.519
COMET-MQM_2021	0.490
C-SPEC	0.484
COMET-QE-DA_2021-src	0.484
bleurt-21-beta	0.483
bleurt-20	0.483
COMETinho-DA	0.457
C-SPECpn	0.454
Prism	0.440
YiSi-1	0.425
BERTScore	0.417
COMET-QE-MQM_2021-src	0.379
chrF	0.371
regEMT	0.369
COMETinho-MQM	0.348
OpenKiwi-MQM-src	0.333
YiSi-2-src	0.229
regEMT-baseline	0.079
regEMT-src	0.065
regEMT-baseline-src	-0.161
TER	-0.791
sentBLEU	-0.881

Table 41: Correlations for English→Japanese: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
bleurt-21-beta	0.978
COMET-DA_2020	0.973
COMET-DA_2021	0.973
bleurt-20	0.972
COMET-MQM_2021	0.972
COMET-QE-DA_2021-src	0.970
COMET-QE-MQM_2021-src	0.969
sentBLEU	0.967
BERTScore	0.964
hLEPOR	0.959
MEE	0.956
MEE2	0.951
YiSi-1	0.948
OpenKiwi-MQM-src	0.948
chrF	0.946
BLEU	0.946
COMETinho-DA	0.944
Prism	0.924
TER	0.903
COMETinho-MQM	0.835
regEMT	0.810
regEMT-baseline	0.377
YiSi-2-src	0.029
regEMT-src	0.005
C-SPECpn	-0.160
regEMT-baseline-src	-0.410
C-SPEC	-0.417

metric	correlation
COMET-DA_2021	0.401
COMET-MQM_2021	0.397
COMET-DA_2020	0.368
COMET-QE-DA_2021-src	0.365
C-SPEC	0.360
C-SPECpn	0.348
bleurt-21-beta	0.340
Prism	0.330
COMET-QE-MQM_2021-src	0.326
bleurt-20	0.323
YiSi-1	0.294
BERTScore	0.255
COMETinho-DA	0.246
OpenKiwi-MQM-src	0.234
MEE2	0.233
chrF	0.201
COMETinho-MQM	0.167
MEE	0.161
hLEPOR	0.157
regEMT	0.122
sentBLEU	0.105
YiSi-2-src	0.051
regEMT-src	0.024
regEMT-baseline	-0.002
TER	-0.078
regEMT-baseline-src	-0.183

Table 42: Correlations for English→Russian: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
COMET-DA_2021	0.946
COMET-DA_2020	0.939
COMET-QE-DA_2021-src	0.927
bleurt-21-beta	0.910
COMET-MQM_2021	0.903
COMETinho-DA	0.900
OpenKiwi-MQM-src	0.892
YiSi-1	0.888
BERTScore	0.851
regEMT	0.823
COMET-QE-MQM_2021-src	0.815
bleurt-20	0.813
Prism	0.750
chrF	0.570
COMETinho-MQM	0.467
YiSi-2-src	0.313
regEMT-src	0.302
C-SPECpn	0.286
C-SPEC	0.235
TER	0.169
regEMT-baseline	0.014
regEMT-baseline-src	-0.039
sentBLEU	-0.156

metric	correlation
COMET-DA_2021	0.270
COMET-QE-DA_2021-src	0.261
COMET-DA_2020	0.247
COMET-MQM_2021	0.246
bleurt-20	0.240
bleurt-21-beta	0.239
C-SPECpn	0.224
C-SPEC	0.224
COMET-QE-MQM_2021-src	0.216
Prism	0.207
COMETinho-DA	0.202
YiSi-1	0.192
BERTScore	0.189
OpenKiwi-MQM-src	0.180
COMETinho-MQM	0.121
regEMT	0.119
YiSi-2-src	0.095
chrF	0.092
regEMT-src	0.016
regEMT-baseline	-0.047
regEMT-baseline-src	-0.187
TER	-0.701
sentBLEU	-0.715

Table 43: Correlations for English→Chinese: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
bleurt-21-beta	0.789
COMET-DA_2020	0.770
bleurt-20	0.768
COMET-MQM_2021	0.763
COMET-DA_2021	0.759
Prism	0.729
COMET-QE-MQM_2021-src	0.712
YiSi-1	0.709
COMET-QE-DA_2021-src	0.708
regEMT	0.702
COMETinho-DA	0.695
BERTScore	0.674
sentBLEU	0.660
chrF	0.647
COMETinho-MQM	0.640
OpenKiwi-MQM-src	0.626
TER	0.615
MTEQA	0.609
C-SPEC	0.191
regEMT-baseline	0.081
regEMT-src	-0.002
regEMT-baseline-src	-0.082
C-SPECpn	-0.267
YiSi-2-src	-0.290

metric	correlation
COMET-DA_2021	0.108
COMET-DA_2020	0.101
regEMT	0.097
COMET-QE-DA_2021-src	0.091
COMET-MQM_2021	0.090
Prism	0.090
bleurt-21-beta	0.081
C-SPEC	0.079
bleurt-20	0.079
YiSi-2-src	0.072
C-SPECpn	0.069
BERTScore	0.068
COMETinho-DA	0.068
OpenKiwi-MQM-src	0.056
chrF	0.054
YiSi-1	0.053
COMET-QE-MQM_2021-src	0.052
COMETinho-MQM	0.052
regEMT-src	0.039
sentBLEU	0.005
regEMT-baseline	-0.081
MTEQA	-0.089
TER	-0.093
regEMT-baseline-src	-0.109

Table 44: Correlations for French→German: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
regEMT	0.964
bleurt-21-beta	0.964
COMET-QE-MQM_2021-src	0.963
bleurt-20	0.963
COMET-QE-DA_2021-src	0.957
COMET-DA_2020	0.955
OpenKiwi-MQM-src	0.953
COMET-MQM_2021	0.949
YiSi-1	0.948
COMET-DA_2021	0.946
chrF	0.941
BERTScore	0.935
COMETinho-DA	0.928
COMETinho-MQM	0.923
TER	0.912
sentBLEU	0.901
regEMT-baseline	0.889
C-SPEC	0.743
YiSi-2-src	0.668
C-SPECpn	0.503
regEMT-baseline-src	0.033
regEMT-src	-0.245

metric	correlation
bleurt-20	0.179
bleurt-21-beta	0.170
C-SPEC	0.157
COMET-DA_2020	0.156
COMET-MQM_2021	0.153
C-SPECpn	0.150
COMET-QE-DA_2021-src	0.146
COMET-DA_2021	0.146
OpenKiwi-MQM-src	0.137
YiSi-1	0.134
COMETinho-DA	0.125
regEMT	0.111
YiSi-2-src	0.110
COMET-QE-MQM_2021-src	0.109
COMETinho-MQM	0.101
BERTScore	0.093
chrF	0.071
sentBLEU	0.070
regEMT-src	-0.027
TER	-0.030
regEMT-baseline	-0.040
regEMT-baseline-src	-0.054

Table 45: Correlations for FLORES Bengali→Hindi: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
Prism	0.990
regEMT	0.987
COMET-QE-DA_2021-src	0.987
COMET-QE-MQM_2021-src	0.986
OpenKiwi-MQM-src	0.982
bleurt-21-beta	0.975
COMETinho-MQM	0.975
COMET-DA_2020	0.974
bleurt-20	0.973
COMET-MQM_2021	0.970
COMET-DA_2021	0.965
COMETinho-DA	0.964
YiSi-1	0.947
BERTScore	0.918
YiSi-2-src	0.898
chrF	0.872
TER	0.871
regEMT-baseline	0.856
sentBLEU	0.784
C-SPECpn	-0.116
C-SPEC	-0.539
regEMT-baseline-src	-0.886
regEMT-src	-0.955

metric	correlation
Prism	0.566
COMET-QE-MQM_2021-src	0.524
COMET-QE-DA_2021-src	0.524
COMET-DA_2020	0.518
COMET-MQM_2021	0.517
COMET-DA_2021	0.510
bleurt-20	0.499
bleurt-21-beta	0.488
C-SPECpn	0.477
YiSi-2-src	0.468
COMETinho-MQM	0.462
COMETinho-DA	0.453
YiSi-1	0.442
C-SPEC	0.418
OpenKiwi-MQM-src	0.412
BERTScore	0.366
chrF	0.327
sentBLEU	0.246
regEMT	0.205
TER	0.108
regEMT-baseline	0.050
regEMT-src	-0.067
regEMT-baseline-src	-0.188

Table 46: Correlations for FLORES Hindi→Bengali: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
COMET-DA_2021	0.999
bleurt-21-beta	0.998
YiSi-1	0.998
chrF	0.998
COMET-MQM_2021	0.997
bleurt-20	0.997
COMET-QE-MQM_2021-src	0.997
COMETinho-DA	0.997
BERTScore	0.995
COMET-DA_2020	0.993
regEMT	0.990
TER	0.981
sentBLEU	0.979
C-SPECpn	0.974
COMETinho-MQM	0.971
OpenKiwi-MQM-src	0.952
C-SPEC	0.942
COMET-QE-DA_2021-src	0.936
regEMT-baseline	0.781
regEMT-src	0.536
YiSi-2-src	0.381
regEMT-baseline-src	0.363

metric	correlation
C-SPEC	0.368
bleurt-20	0.363
bleurt-21-beta	0.359
C-SPECpn	0.340
chrF	0.301
COMET-DA_2021	0.297
COMET-MQM_2021	0.293
YiSi-1	0.293
OpenKiwi-MQM-src	0.286
COMET-QE-DA_2021-src	0.285
COMET-DA_2020	0.281
COMET-QE-MQM_2021-src	0.276
BERTScore	0.270
COMETinho-MQM	0.219
COMETinho-DA	0.209
sentBLEU	0.188
YiSi-2-src	0.153
regEMT-src	0.150
regEMT	0.126
TER	0.074
regEMT-baseline	-0.014
regEMT-baseline-src	-0.053

Table 47: Correlations for FLORES Xhosa→Zulu: system-level Pearson (left panel) , segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.

metric	correlation
bleurt-21-beta	1.000
chrF	0.999
YiSi-1	0.998
bleurt-20	0.998
BERTScore	0.997
COMET-MQM_2021	0.997
COMETinho-DA	0.996
COMET-DA_2021	0.996
COMETinho-MQM	0.991
COMET-QE-MQM_2021-src	0.990
COMET-DA_2020	0.990
regEMT	0.983
TER	0.978
OpenKiwi-MQM-src	0.973
COMET-QE-DA_2021-src	0.953
sentBLEU	0.903
YiSi-2-src	0.758
C-SPECpn	0.713
regEMT-baseline	0.681
C-SPEC	0.604
regEMT-baseline-src	0.432
regEMT-src	-0.044

metric	correlation
COMET-DA_2021	0.571
bleurt-20	0.564
bleurt-21-beta	0.559
C-SPEC	0.552
C-SPECpn	0.552
COMET-MQM_2021	0.550
COMET-DA_2020	0.545
YiSi-1	0.544
COMET-QE-MQM_2021-src	0.538
chrF	0.530
COMET-QE-DA_2021-src	0.530
OpenKiwi-MQM-src	0.523
BERTScore	0.491
YiSi-2-src	0.472
COMETinho-DA	0.436
COMETinho-MQM	0.423
sentBLEU	0.381
TER	0.296
regEMT	0.202
regEMT-baseline	0.022
regEMT-src	-0.010
regEMT-baseline-src	-0.037

Table 48: Correlations for FLORES Zulu→Xhosa: system-level Pearson (left panel), segment-level Kendall-Like (right panel). Primary submissions are bolded, and baselines are underlined.