# Results of WMT23 Metrics Shared Task: Metrics might be Guilty but References are not Innocent

**Markus Freitag**[(1)], **Nitika Mathur**[(2)], **Chi-kiu Lo 羅致翹**[(3)], **Eleftherios Avramidis**[(4)], **Ricardo Rei**[(5,6,7)], **Brian Thompson**[(8)], **Tom Kocmi**[(9)], **Frédéric Blain**[(10)], **Daniel Deutsch**[(1)], **Craig Stewart**[(11)], **Chrysoula Zerva**[(7,12)], **Sheila Castilho**[(13)], **Alon Lavie**[(11)], **George Foster**[(1)]
[(1)]Google Research [(2)]Oracle Digital Assistant [(3)]National Research Council Canada
[(4)]German Research Center for Artificial Intelligence (DFKI) [(5)]Unbabel [(6)]INESC-ID
[(7)]Instituto Superior Técnico [(8)]AWS AI Labs [(9)]Microsoft [(10)]Tilburg University [(11)]Phrase
[(12)]Instituto de Telecomunicações [(13)]Dublin City University
wmt-metrics@googlegroups.com

## Abstract

This paper presents the results of the WMT23 Metrics Shared Task. Participants submitting automatic MT evaluation metrics were asked to score the outputs of the translation systems competing in the WMT23 News Translation Task. All metrics were evaluated on how well they correlate with human ratings at the system and segment level. Similar to last year, we acquired our own human ratings based on expert-based human evaluation via Multidimensional Quality Metrics (MQM). Following last year's success, we also included a challenge set subtask, where participants had to create contrastive test suites for evaluating metrics' ability to capture and penalise specific types of translation errors. Furthermore, we improved our meta-evaluation procedure by considering fewer tasks and calculating a global score by weighted averaging across the various tasks.

We present an extensive analysis on how well metrics perform on three language pairs: Chinese→English, Hebrew→English on the sentence-level and English→German on the paragraph-level. The results strongly confirm the results reported last year, that neural-based metrics are significantly better than non-neural metrics in their levels of correlation with human judgments. Further, we investigate the impact of bad reference translations on the correlations of metrics with human judgment. We present a novel approach for generating synthetic reference translations based on the collection of MT system outputs and their corresponding MQM ratings, which has the potential to mitigate bad reference issues we observed this year for some language pairs. Finally, we also study the connections between the magnitude of metric differences and their expected significance in human evaluation, which should help the community to better understand and adopt new metrics.

| Metric | | avg corr |
|---|---|---|
| XCOMET-Ensemble | **1** | **0.825** |
| XCOMET-QE-Ensemble* | 2 | 0.808 |
| MetricX-23 | 2 | 0.808 |
| GEMBA-MQM* | 2 | 0.802 |
| MetricX-23-QE* | 2 | 0.800 |
| mbr-metricx-qe* | 3 | 0.788 |
| MaTESe | 3 | 0.782 |
| CometKiwi* | 3 | 0.782 |
| COMET | 3 | 0.779 |
| BLEURT-20 | 3 | 0.776 |
| KG-BERTScore* | 3 | 0.774 |
| sescoreX | 3 | 0.772 |
| cometoid22-wmt22* | 4 | 0.772 |
| docWMT22CometDA | 4 | 0.768 |
| docWMT22CometKiwiDA* | 4 | 0.767 |
| Calibri-COMET22 | 4 | 0.767 |
| Calibri-COMET22-QE* | 4 | 0.755 |
| YiSi-1 | 4 | 0.754 |
| MS-COMET-QE-22* | 5 | 0.744 |
| prismRef | 5 | 0.744 |
| mre-score-labse-regular | 5 | 0.743 |
| BERTscore | 5 | 0.742 |
| XLsim | 6 | 0.719 |
| f200spBLEU | 7 | 0.704 |
| MEE4 | 7 | 0.704 |
| tokengram_F | 7 | 0.703 |
| embed_llama | 7 | 0.701 |
| BLEU | 7 | 0.696 |
| chrF | 7 | 0.694 |
| eBLEU | 7 | 0.692 |
| Random-sysname* | 8 | 0.529 |
| prismSrc* | 9 | 0.455 |

Table 1: Official ranking of primary submissions to the WMT23 Metric Task. The final score is the weighted average correlation over 10 different tasks. Starred metrics are reference-free, and underlined metrics are baselines. See Table 18 for the pairwise comparisons from which the ranks were derived.

## 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 drive the development of new metrics. We eval-

[1] https://wmt-metrics-task.github.io/



*Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 578–628
December 6–7, 2023. 

uate reference-based automatic metrics that score MT output by comparing the translations 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 and sentence level. The final ranking of this year's submitted primary metrics is shown in Table 1. Below are some key details and changes for this year's metric shared task:

- **Language Pairs**: For this year, we focus on three main language pairs: (i) One language pair with paragraph-level test sets: English→German (en→de), (i) one low-resource language pair with sentence-level test sets: Hebrew→English (he→en), (iii) one high-resource language pair with sentence-level test sets: Chinese→English (zh→en).

- **Human Evaluation**: Like last year, we collected our own human ratings for our three language pairs from professional translators via MQM (Lommel et al., 2014; Freitag et al., 2021). We released and uploaded[2] all MQM annotations, and we recommend using Marot[3] for looking into this data.

- **Meta Evaluation**: This year's meta-evaluation is significantly streamlined from last year's. Instead of 201 tasks, we use just 10, designed to capture complementary ranking and linearity properties at system- and segment-level granularity. We replace Kendall's tau at the segment level with a version of pairwise accuracy that gives metrics credit for correctly predicting ties in human scores, while automatically calibrating for each metric's natural scale (Deutsch et al., 2023). Instead of averaging per-task ranks to derive an overall score for each metric, we simply average correlation/accuracy scores across tasks. This places metric scores on an absolute scale, and makes them independent of the performance of other metrics. Finally, we compute top-level significance clusters to provide a clearer global ranking of participating metrics.

- **Synthetic Reference**: The MQM scores for the human reference translation for zh→en were unexpectedly low, ranking humans below almost all WMT submissions. We investigate the impact of bad reference translations on reference-based metrics and propose a novel approach to create a synthetic reference translation from all WMT submissions and their corresponding MQM scores.

- **Challenge Sets Subtask**: For the second year, we include a decentralized sub-task on challenge sets, in which test sets are submitted by different research teams targeting to reveal metrics' abilities or the weaknesses in evaluating particular translation phenomena. We received three challenge sets covering a wide range of translation errors and linguistic phenomena in more than a hundred translation directions.

- **Understand Magnitude of Score Difference**: This year, we include two analyses to understand the meaning of the score differences that metrics present with respect to the statistical significance of MT system rankings according to human annotations and metric scores. These analyses provide additional assistance for MT researchers to build an intuition on the relationship between the magnitude of metric score differences and the reliability of the improved translation quality.

- **MTME**: Similar to last year, all the data has been uploaded to MTME[4], and all results in this paper are calculated with this analysis tool. We encourage every metric developer to use MTME to calculate contrastive scores to enhance consistency and comparability going forward.

Our main findings are:

- XCOMET-Ensemble is the winner of the WMT23 Metrics Shared Task (Table 1).

- High correlations between automatic metrics and human judgments at the segment level do not necessarily guarantee high correlations at the system level (Figure 5).

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

[3] https://github.com/google-research/google-research/tree/master/marot

[4] https://github.com/google-research/mt-metrics-eval

- Reference quality matters: The low quality reference for zh→en significantly lowered the correlation of all metrics with human judgement (Section 8).
- We determined the magnitude of score differences required to produce a statistically significant difference in human judgment for each metric, revealing that even minor score differences of the top performing metrics can be statistical significant with high probability (Section 7).
- Results from the challenge sets independently agreed with our findings that the quality of reference matters. Developing reference-free metrics is worth further exploration, and metric researchers are advised to investigate into the influence of language-agnostic multilingual embeddings on MT evaluation. It is equally important for metric researchers to test the performance of metrics in diverse collection of linguistic phenomena and wider landscape of translation quality in order to minimize unexpected behaviours of metrics (Section 10).

The rest of the paper is organized as follows: Section 2 describes the test data and additional MT systems that we trained. Section 3 presents an overview of the conducted expert-based human evaluation. Section 4 describes the metrics evaluated this year (baselines and participants). Section 5 describes the conducted meta-evaluation. Section 6 reports our main results. Section 7 interprets and evaluates metrics' scores beyond correlations. Section 8 analyses the impact of bad reference translations on the various metrics. Section 9 summarizes our results for additional WMT23 Translation task language-pairs based on their Direct Assessment human evaluation. Section 10 presents a description of the submitted challenge sets along with their findings. Finally, Section 11 presents our most relevant conclusions.

## 2 Translation Systems

Similar to previous years' editions, the source, reference texts, and MT system outputs for the metrics task are mainly derived from the WMT23 General MT Shared Task. In addition to the MT system outputs from the WMT evaluation campaign, we included translations from two additional MT systems which we deemed interesting for evaluation.

### 2.1 WMT Test Sets

We use test sets prepared by the WMT23 General MT Shared Task (Kocmi et al., 2023). For our three main language pairs, the test sets contain 557 en→de, 1910 he→en, and 1976 zh→en segments. This year, the test sets cover up to five domains from the following list: news, conversational, user reviews, manuals, and social. Each language pair contains a comparable number of sentences from each domain, resulting in reasonably balanced test sets.

English→German contains four balanced domains: news, social, conversational, and user reviews. In contrast to other language pairs, segments are paragraphs rather than sentences.

Hebrew→English contains only news and user reviews domains. This language pair has two human references, but one of them (refA) is suspected of being a post-edited Online-B system output.

Chinese→English contains news, user reviews, and manuals. The first two domains contain around 750 sentences, while manuals contains around 500.

The reference translations provided for the test sets are produced by professional translators.

For more details regarding the news test sets, we refer the reader to the WMT23 General MT Shared Task findings paper (Kocmi et al., 2023).

### 2.2 Additional MT Output

Similar to last year, we made an effort to expand the pool of translations beyond the WMT submissions, which can potentially be quite similar to each other. We added translations which we expected to differ in two main ways from the submissions: 1) by using a massively multilingual model; and 2) by generating with MBR decoding;

For our multilingual model, we selected the 3.3B parameter NLLB200 model (NLLB Team et al., 2022) via the huggingface (Wolf et al., 2020) interface. We found NLLB200 to significantly outperform the M2M100 (Fan et al., 2021) that we used last year.

Minimum Bayes Risk (MBR) decoding has recently gained attention in MT as a decision rule, with the potential to overcome some of the biases of MAP decoding in NMT (Eikema and Aziz, 2020; Müller and Sennrich, 2021; Eikema and Aziz, 2021; Freitag et al., 2022; Fernandes et al., 2022). MBR decoding centrally relies on a reference-based utility metric: its goal is to identify a hypothesis with a high estimated utility (expectation

under model distribution) with the hope that a high estimated utility translates into a high actual utility (with respect to a human reference). In practice, this means generating several candidate translations and finding the translation that is most similar to the rest of the candidate translations.

We produced both the top-1 greedy translation and MBR outputs. For MBR, we sampled 100 translation candidates from the model via Epsilon sampling (Hewitt et al., 2022; Freitag et al., 2023). We used `epsilon_cutoff=0.02` and `eta_cutoff=0.0`. This year, we used sentence-level BLEU from sacreBLEU (Post, 2018) with the default 'a13' tokenizer and the 'floor' smoothing method as utility function only.

## 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 crowdsourced 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. Similar to last year, we decided to conduct our own MQM-based human evaluation on a subset of submissions and language pairs that are most interesting for evaluating current 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 (Hebrew→English). The error annotation typology and guidelines used by Google's and Unbabel's annotators differ slightly and are 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 18 professional translators (10 for English→German, 8 for Chinese→English) having access to the full document context. Each segment gets annotated by a single rater. Instead of assigning a scalar value to each translation, annotators were instructed to label error spans within each segment in a document, paying particular attention to document context. Each error was highlighted in the text, and labelled with an error category and a severity. 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. 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 previous two metrics tasks.

| 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 Hebrew→English

The annotations for the Hebrew→English language pair were sourced from Unbabel, who engaged four professional native language annotators possessing extensive translation experience. Much like Google's approach, these annotators were provided with the full document context, comprising up to ten segments. Their task was to identify and classify errors by highlighting them, following Unbabel's MQM 3.0 typology[5].

The annotators were instructed to classify the errors based on severity, with Unbabel's classification encompassing not only "Minor" and "Major" error severities (analogous to Google's criteria) but also a "Critical" error severity. However, to ensure consistency in our evaluation process, we opted to align with the Google methodology outlined previously. Specifically, we treated all annotated "Critical" errors as "Major" errors, and we applied a weighting scheme for punctuation errors, as detailed in Table 2.

[5]see Unbabel Annotation Guidelines - Typology 3.0

### 3.3 Human Evaluation Results

Due to the fact that we ran our own human evaluation, we were only able to evaluate a subset of the test segments. In Table 3, you can see the number of segments and documents for each language pair and test set that we used for human evaluation. We followed a simple and consistent approach to downsample the data: we considered each document, while only keeping the first 10 sentences of each document. By doing this, we did not need to discard most of the documents and only needed to crop longer documents. The English→German test is on the paragraph-level, and we had to discard two documents as the first paragraph already contained more than 10 sentences. In all cases, the MQM score for a segment is the sum of the scores for the errors in that segment, and the MQM score for a test set is the average of the MQM scores of the segments that were annotated.

The results of the MQM human evaluation can be seen in Table 4. Most of the reference translations are ranked first, except for refA for Chinese→English. Not ranking the human evaluation on top of the MT output is usually a signal for a corrupt human evaluation. We double-checked the annotation for refA and can confirm that the reference translation indeed contained many errors.

## 4 Baselines and Submissions

We computed scores for several baseline metrics in order to compare submissions against previous well-studied metrics. We will start by describing those baselines, and then we will describe the submissions from participating teams. An overview of the evaluated metrics can be seen in Table 5.

### 4.1 Baselines

**SacreBLEU baselines** We use the following metrics from SacreBLEU (Post, 2018) as baselines:

- **BLEU (Papineni et al., 2002)** is based on the precision of $n$-grams between the MT output and its reference weighted by a brevity penalty. Using SacreBLEU we obtained sentenceBLEU values using the `sentence_bleu` Python function and for corpus-level BLEU we used `corpus_bleu` (both with default arguments[6]).

- **F200SPBLEU (NLLB Team et al., 2022)** are BLEU scores computed with subword tokenization done by the standardized FLORES-200 Sentencepiece models. We used the command line SacreBLEU to compute the sentence level F200SPBLEU[7] and we average the segment-level scores to obtain a corpus-level score.

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

**BLEURT (Sellam et al., 2020)** is a learned metric fine-tuned on Direct Assessments (DA). Unlike COMET, BLEURT encodes the translation and the reference together and utilizes the `[CLS]` token as an embedding to represent the pair. We employed the BLEURT20 checkpoint (Pu et al., 2021), which was trained on top of RemBERT using DA data from previous shared tasks spanning from 2015 to 2019, along with additional synthetic data created from Wikipedia articles.

**COMET (Rei et al., 2022a)** is a learned metric fine-tuned using DA from previous WMT Translation shared tasks. This metric relies on sentence embeddings from the source, translation, and reference to produce a final score. We utilized the default model `wmt22-comet-da` provided in version 2.0.2 of the `Unbabel/COMET` framework. This model employs XLM-R large as its backbone model and is trained on data from the 2017 to 2019 WMT shared tasks, in combination with the MLQE-PE corpus (Fomicheva et al., 2022).

**COMETKIWI (Rei et al., 2022b)** is a reference-free learned metric that functions similarly to

[6] lnrefs.1|case.mixed|lang.LANGPAIR|tok.13a|smooth.exp| version.2.3.0

[7] nrefs:1|case:mixed|eff:yes|tok:flores200|smooth:exp| version:2.3.0

[8] chrF2|lang.LANGPAIR|nchars.6|space.false|version.2.3.0

| language | news | social | speech | user reviews | manuals |
|---|---|---|---|---|---|
| en→de | 104/139 (30/30) | 206/212 (79/79) | 58/113 (23/25) | 92/93 (58/58) | n/a |
| he→en | 619/1558 (68/70) | n/a | n/a | 201/352 (26/26) | n/a |
| zh→en | 377/763 (38/38) | n/a | n/a | 677/726 (127/127) | 123/487 (14/14) |

Table 3: Numbers of MQM-annotated segments per domain (number of docs in brackets).

| English→German ↓ System | all | news | social | speech | user-reviews |
|---|---|---|---|---|---|
| refA | 2.96 | 3.12 | 2.02 | 4.74 | 3.77 |
| GPT4-5shot | 3.72 | 4.00 | 2.41 | 6.51 | 4.60 |
| ONLINE-W | 3.95 | 2.69 | 2.62 | 5.90 | 7.13 |
| ONLINE-B | 4.71 | 4.35 | 3.14 | 5.96 | 7.85 |
| ONLINE-Y | 5.64 | 4.45 | 3.67 | 7.48 | 10.26 |
| ONLINE-A | 5.67 | 4.40 | 3.84 | 7.78 | 9.87 |
| ONLINE-G | 6.57 | 6.43 | 4.12 | 7.93 | 11.38 |
| ONLINE-M | 6.94 | 4.87 | 4.41 | 8.30 | 14.08 |
| Lan-BridgeMT | 8.67 | 7.99 | 5.55 | 9.72 | 15.78 |
| LanguageX | 9.25 | 8.43 | 5.74 | 14.23 | 14.92 |
| NLLB_Greedy | 9.54 | 8.29 | 5.20 | 14.82 | 17.35 |
| NLLB_MBR_BLEU | 10.79 | 9.93 | 5.53 | 17.75 | 19.18 |
| AIRC | 14.23 | 14.32 | 8.34 | 20.34 | 23.45 |

| Hebrew→English ↓ System | all | news | user-reviews |
|---|---|---|---|
| refA | 1.17 | 1.28 | 0.86 |
| GPT4-5shot | 1.33 | 1.29 | 1.48 |
| ONLINE-A | 1.38 | 1.34 | 1.50 |
| ONLINE-B | 1.55 | 1.60 | 1.39 |
| GTCOM_DLUT | 1.89 | 1.85 | 1.99 |
| UvA-LTL | 1.92 | 1.80 | 2.30 |
| ONLINE-G | 2.06 | 2.06 | 2.04 |
| ONLINE-Y | 2.35 | 2.42 | 2.12 |
| LanguageX | 2.38 | 2.33 | 2.53 |
| Samsung_Research_Philippines | 3.23 | 3.62 | 2.05 |
| NLLB_MBR_BLEU | 3.68 | 3.83 | 3.20 |
| NLLB_Greedy | 3.79 | 3.98 | 3.19 |
| Lan-BridgeMT | 3.79 | 3.81 | 3.74 |

| Chinese→English ↓ System | all | news | manuals | user-reviews |
|---|---|---|---|---|
| Lan-BridgeMT | 2.10 | 2.31 | 1.28 | 2.13 |
| GPT4-5shot | 2.31 | 2.26 | 2.01 | 2.39 |
| Yishu | 3.23 | 3.34 | 1.67 | 3.46 |
| ONLINE-B | 3.39 | 3.27 | 1.78 | 3.74 |
| HW-TSC | 3.40 | 3.40 | 1.83 | 3.68 |
| ONLINE-A | 3.79 | 2.90 | 1.83 | 4.63 |
| ONLINE-Y | 3.79 | 3.47 | 2.84 | 4.14 |
| ONLINE-G | 3.86 | 3.58 | 2.02 | 4.34 |
| ONLINE-W | 4.06 | 3.84 | 2.16 | 4.53 |
| LanguageX | 4.23 | 4.05 | 2.84 | 4.59 |
| IOL_Research | 4.59 | 3.60 | 1.85 | 5.63 |
| refA | 4.83 | 5.04 | 5.17 | 4.65 |
| ONLINE-M | 5.43 | 4.71 | 2.98 | 6.28 |
| ANVITA | 6.08 | 5.17 | 2.97 | 7.15 |
| NLLB_MBR_BLEU | 6.36 | 6.57 | 3.39 | 6.78 |
| NLLB_Greedy | 6.57 | 6.70 | 2.95 | 7.16 |

Table 4: MQM human evaluations for generalMT2023. Lower average error counts represent higher MT quality. Systems above any solid line are significantly better than those below, based on all domains with $p < 0.05$.

BLEURT, but instead of encoding the translation along with its reference, it uses the source. We utilized the `wmt22-cometkiwi-da` model, which was a top-performing reference-free metric from last year's shared task. This reference-free metric is fine-tuned on the same data as `wmt22-comet-da` using the version 2.0.2 of the `Unbabel/COMET` framework.

**DOCWMT22COMETDA (Vernikos et al., 2022)** is the document-level version of `wmt22-comet-da`, which computes the BERT embeddings using multi-sentence context instead of just the single sentence.

DOCWMT22COMETKIWIDA is the document-level version of WMT22-COMETKIWI-DA (QE) which computes the BERT embeddings using multi-sentence context instead of just the single sentence.

**MS-COMET-QE-22 (Kocmi et al., 2022b)** is built on top of COMET by Microsoft Research using proprietary data. This metric is trained on a several times larger set of human judgements compared to COMET-baseline, covering 113 languages and 15 domains. Furthermore, the authors propose filtering of human judgement with potentially low quality to further improve the model. The metric calculated scores in quality estimation fashion with only source segment and MT hypothesis.

**PRISMREF and PRISMSRC (Thompson and Post, 2020a,b)** PRISMREF is the reference-based PRISM that uses a multilingual MT model in zero-shot paraphrase model to score the candidate translation conditioned on the reference sentence, and the reference sentence conditioned on the candidate translation, and averages the two scores. PRISM-SRC is the source-based (i.e. QE as a metric) PRISM that uses a multilingual MT model to force-decode and score the candidate translation conditioned on the source sentence.

**RANDOM-SYSNAME** is a random metric that takes the system name as the only parameter. For each translation system, the metric computes the mean value X as $sha256(sysname)[0]\%10$. It uses discrete scores. Segment-level scores follow

| | metric | broad category | supervised | ref. free | citation | availability (https://github.com/) |
|---|---|---|---|---|---|---|
| baselines | BLEU | lexical overlap | | | Papineni et al. (2002) | mjpost/sacrebleu |
| | F200SPBLEU | lexical overlap | | | NLLB Team et al. (2022) | mjpost/sacrebleu |
| | CHRF | lexical overlap | | | Popović (2015) | mjpost/sacrebleu |
| | BERTSCORE | embedding similarity | | | Zhang et al. (2020) | Tiiiger/bert_score |
| | BLEURT | fine-tuned metric | ✓ | | Sellam et al. (2020) | google-research/bleurt |
| | COMET | fine-tuned metric | ✓ | | Rei et al. (2022a) | Unbabel/COMET |
| | COMETKIWI | fine-tuned metric | ✓ | ✓ | Rei et al. (2022b) | Unbabel/COMET |
| | DOCWMT22COMETDA | fine-tuned metric | ✓ | | Vernikos et al. (2022) | amazon-research/doc-mt-metrics |
| | DOCWMT22COMETKIWIDA | fine-tuned metric | ✓ | ✓ | Vernikos et al. (2022) | amazon-research/doc-mt-metrics |
| | MS-COMET-QE-22 | fine-tuned metric | ✓ | ✓ | Kocmi et al. (2022b) | MicrosoftTranslator/MS-Comet |
| | PRISMREF | MT-model metric | ✓ | | Thompson and Post (2020a,b) | thompsonb/prism |
| | PRISMSRC | MT-model metric | ✓ | ✓ | Thompson and Post (2020a,b) | thompsonb/prism |
| | RANDOM-SYSNAME | random baseline | | ✓ | — | (not available) |
| | YISI-1 | embedding similarity | | | Lo (2019) | chikiulo/yisi |
| primary submissions | CALIBRI-COMET22 | fine-tuned metric | ✓ | | — | (not available) |
| | CALIBRI-COMET22-QE | fine-tuned metric | ✓ | ✓ | — | (not available) |
| | COMETOID22-WMT22 | fine-tuned metric | ✓ | ✓ | Gowda et al. (2023) | marian-nmt/marian-dev |
| | EBLEU | embedding similarity | | | ElNokrashy and Kocmi (2023) | munael/ebleu-mt-metrics-wmt23 |
| | EMBED_LLAMA | embedding similarity | | | Dreano et al. (2023a) | SorenDreano/embed_llama |
| | GEMBA-MQM | LLM prompt-based metric | | ✓ | Kocmi and Federmann (2023) | MicrosoftTranslator/GEMBA |
| | KG-BERTSCORE | embedding similarity | | ✓ | Wu et al. (2023) | (not available) |
| | MATESE | fine-tuned metric | ✓ | | Perrella et al. (2022) | SapienzaNLP/MaTESe |
| | MBR-METRICX-QE | fine-tuned metric | ✓ | ✓ | Naskar et al. (2023) | (not available) |
| | MEE4 | lexical & embedding similarity | | | Mukherjee and Shrivastava (2023) | AnanyaCoder/WMT22Submission |
| | METRICX-23 | fine-tuned metric | ✓ | | Juraska et al. (2023) | google-research/metricx |
| | METRICX-23-QE | fine-tuned metric | ✓ | ✓ | Juraska et al. (2023) | google-research/metricx |
| | MRE-SCORE-LABSE-REGULAR | fine-tuned metric | ✓ | | Viskov et al. (2023) | NL2G/efficient-llm-metrics |
| | SESCORE | fine-tuned metric | ✓ | | Xu et al. (2022) | xu1998hz/SEScore |
| | TOKENGRAM_F | lexical overlap | | | Dreano et al. (2023b) | SorenDreano/tokengram_F |
| | XCOMET-ENSEMBLE | fine-tuned metric | ✓ | | Guerreiro et al. (2023) | Unbabel/COMET |
| | XCOMET-QE-ENSEMBLE | fine-tuned metric | ✓ | ✓ | Guerreiro et al. (2023) | Unbabel/COMET |
| | XLSIM | fine-tuned metric | ✓ | | Mukherjee and Shrivastava (2023) | AnanyaCoder/XLsim |

Table 5: Baseline metrics and primary submissions for the metrics task. We categorize metrics into 3 major classes: lexical, embedding similarity and fine-tuned metrics. Regarding fine-tuned metrics we have metrics that use human quality scores such as DA or MQM and metrics that use synthetic labels for fine-tuning (3rd column).

Gaussian distribution around mean value X (in a range 0-9) and standard deviation of 2.

**YiSi-1 (Lo, 2019)** is a MT evaluation metric that 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 pre-trained language models (e.g. RoBERTa, CamemBERT, XLM-RoBERTa, etc.).

### 4.2 Metric Submissions

The rest of this section summarizes the participating metrics.

**CALIBRI-COMET22 and CALIBRI-COMET22-QE** apply a post-processing approach to ratings provided by COMET. It uses `Unbabel/wmt22-comet-da` as the backbone for the referenced CALIBRI-COMET22 and `Unbabel/wmt22-cometkiwi-da` as the backbone for the unreferenced CALIBRI-COMET22-QE metric. The information whether a translation is error-free from MQM ratings (e.g. under Google's MQM error weighting, error-free translations have a score of 0) can be recovered. It then aims to calibrate the scores of the backbone model with respect to this binary error-freeness label using isotonic regression. During test time, it takes the samples for a given tuple (lang-pair, test-set, domain, ref, system-id) and employs a heuristic strategy to select samples from previous years that match the test sample score distribution. It then fits an isotonic regression model to the selected samples and transforms the test scores accordingly. The main idea is that in this way, the averaged system-level score can be interpreted as the fraction of error-free translations.

**COMETOID22 (Gowda et al., 2023)** is a reference-free metric created using knowledge distillation from reference-based metrics. Using COMET-22 as a teacher metric, it scores the MT outputs submitted to the WMT News/General Machine Translation task since 2009. A student metric, called COMETOID22, is then trained to mimic the teacher scores without using reference translation. The student metric has the same architecture as COMET-QE, and is initialized with pretrained weights from InfoXLM, a multilingual language model. We submit three variants: COMETOID22-WMT{21,22,23}, where the suffix indicates the training data cut-off year.

**COMETKIWI XL/XXL (Rei et al., 2023)** shares the same architecture as the COMETKIWI baseline but replaces InfoXLM with XLM-R XL (3.5B) and XXL (10.7B). In terms of training data, these models are trained on the same dataset as COMETKIWI, along with newly released Direct Assessments (DA) for Indian languages, which were introduced as additional training data for this year's Quality Estimation (QE) shared task (Blain et al., 2023).

**EBLEU (ElNokrashy and Kocmi, 2023)** String-based metrics such as BLEU and CHRF depend on string similarity as proxy for meaning similarity between candidate and target sentences. EBLEU stands for 'Embedded BLEU' and is loosely inspired by it. In EBLEU, we match candidate and target tokens approximately using non-contextual word embeddings and a word-to-word similarity map in a form we have dubbed "relative meaning diffusion tensors".

**EMBED_LLAMA (Dreano et al., 2023a)** relies on pretrained Llama2 embeddings, without any fine-tuning, to transform sentences into a vector space that establishes connections between geometric and semantic proximities. This metrics draws inspiration from Word2vec, and utilizes cosine distance for the purpose of estimating similarity or dissimilarity between sentences.

**GEMBA-MQM (Kocmi and Federmann, 2023)** is a LLM-enabled metric for error quality span marking. It uses three-shot prompting with the GPT4 model. In contrast to EAPrompt (Lu et al., 2023), it does not require language specific examples and requires only a single prompt.

**HWTSC-EE-METRIC and KG-BERTSCORE (Wu et al., 2023)** EE stands for Entropy Enhanced MT Metrics and aims at achieving a more balanced system-level rating by assigning weights to segment-level scores produced by MT metrics. The weights are determined by the difficulty of a segment determined by the entropy between the hypothesis-reference pair. This year, the COMET metric is utilized as the backbone of our EE metrics. The model we use is WMT22-COMET-DA.

KG-BERTSCORE incorporates multilingual knowledge graph into BERTSCORE and generates the final evaluation score by linearly combining the results of KGSCORE and BERTSCORE, in which

we use COMET-QE to calculate BERTSCORE this year.

**MATESE (Perrella et al., 2022)** leverages transformer-based encoders to identify error spans in translations, and classify their severity between Minor and Major. Differently from last year's version, MATESE is now based on DeBERTa for evaluating translations towards English, and InfoXLM for German and Russian. Furthermore, it has been re-trained using also the MQM data released at WMT2022.

**MBR-METRICX-QE (Naskar et al., 2023)** MBR decoding with neural utility metrics like BLEURT is known to be effective in generating high quality machine translations. We use the underlying technique of MBR decoding and develop an MBR based reference-free quality estimation metric. Our method uses an evaluator machine translation system and a reference-based utility metric (specifically BLEURT and METRICX) to calculate a quality estimation score of a model. We report results related to comparing different MBR configurations and utility metrics.

**MEE4 (Mukherjee and Shrivastava, 2023)** is an unsupervised, reference-based metric (an improved version of MEE) focusing on computing contextual and syntactic equivalences, along with lexical, morphological, and semantic similarity. The goal is to comprehensively evaluate the fluency and adequacy of MT outputs while also considering the surrounding context. Fluency is determined by analysing syntactic correlations, while context is evaluated by comparing sentence similarities using sentence embeddings. The ultimate score is derived from a weighted amalgamation of three distinct similarity measures: a) Syntactic similarity, which is established using a modified BLEU score. b) Lexical, morphological, and semantic similarity, quantified through explicit unigram matching. c) Contextual similarity, gauged by sentence similarity scores obtained from the Language-Agnostic BERT model.

**METRICX-23 and METRICX-23-QE (Juraska et al., 2023)** are learned reference-based and reference-free (respectively) regression metrics based on the mT5 encoder-decoder language model. They further fine-tune the mT5-XXL checkpoint on direct assessment data from 2015-2020 and MQM data from 2020 to 2021 as well as synthetic data. There are two contrastive submissions, "b" and "c", for both the reference-based and QE metrics. The "b" variant additionally trains on MQM data from 2022 and the "c" variant uses the PaLM-2 language model (Anil et al., 2023) to initialize the metric instead of mT5.

**MRE-SCORE (Viskov et al., 2023)** is a trained metric that is based on the encoder part of mT0-large model. We use a concatenation of source, reference and hypothesis texts for input. Additionally, some of the variants of the model uses contextual embeddings from LaBSE.

**SESCOREX (Xu et al., 2023b) and INSTRUCTSCORE (Xu et al., 2023c)** SESCOREX is an improved version of SESCORE2 (Xu et al., 2023a). Building upon the established strengths of SESCORE2, we utilize its framework for synthetic data generation to pre-train our scoring model. To further elevate the performance of SESCOREX, we introduce two key modifications: fine-tuning human rating data and transitioning the scoring backbone model to the MT5-xl model. INSTRUCTSCORE is an open-source, explainable evaluation metric for text generation. Utilizing explicit human guidelines and GPT4's implicit knowledge, we fine-tune an Llama model to provide evaluation metrics along with diagnostic reports that align with human assessments. Unlike traditional neural metrics, INSTRUCTSCORE evaluates text generation by providing a quality score based on detailed error explanations.

**SLIDE (Raunak et al., 2023)** Building metrics explicitly for document-level MT quality estimation has been challenging owing to the lack of large-scale document-level human annotated datasets. In this submission, we present a metric named SLIDE (Sliding Document Evaluator), which operates at the span of multiple sentences or paragraphs by way of an overlapping sliding window. SLIDE feeds each chunk into a source-based COMET model, with scores over overlapping chunks accumulated to produce a system-level score. SLIDE is motivated by two ideas: (1) Since COMET's underlying encoder is trained on wider contexts, we might observe generalizable evaluation behaviour beyond typical sentences-level lengths, within certain length limits and (2) since a sentence's evaluation will differ at different positions within a document, it may be helpful to evaluate each sentence in multiple different contexts.

**TOKENGRAM_F (Dreano et al., 2023b)** is an F-score-based evaluation metric for machine translation that is heavily inspired by CHRF++. By replacing word-grams with token-grams obtained from contemporary tokenization algorithms, TOKENGRAM_F captures similarities between words sharing the same semantic roots and thus obtains more accurate ratings.

**XCOMET-XL/XXL (Guerreiro et al., 2023)** is a new COMET (Rei et al., 2020) model that is designed to identify error spans in sentences and generate a final quality score, making it a more interpretable learnt metric. This metric is optimized for both regression and sequence tagging, and it can be used with or without references. XCOMET-QE submission results from the same model but running inference without a reference. These models utilize XLM-R XL or XXL as their backbone models, with XCOMET-XL having 3.5B parameters and XCOMET-XXL having 10.7B parameters. The training process for this metric occurs in stages, starting with DAs and then is fine-tuning on MQM data. XCOMET-ENSEMBLE is an ensemble between 1 XL and 2 XXL checkpoints that result from the different training stages.

**XLSIM (Mukherjee and Shrivastava, 2023)** is a supervised reference-based metric that regresses on human scores provided by WMT (2017-2022). Using a cross-lingual language model XLM-RoBERTa, we train a supervised model using a Siamese network architecture with cosine similarity loss.

## 5 Meta Evaluation

Our main goal in evaluating metrics is to establish a ranking that reflects a metric's performance across a range of settings and applications. Combining results from different settings is challenging because correlations with human gold scores have different ranges and may be subject to differing degrees of noise. There are also many ways of measuring correlation, with different strengths and weaknesses, and it is often not clear which is best in a given setting.

Last year, our approach was to define a large number of "tasks" (201 in total) that varied along dimensions such as language pair, domain, granularity, correlation statistic, etc. For each task, we used pairwise significance tests to establish a dense clustered ranking of participating metrics (e.g., 1, 1, 1, 2, 3, 3, ...). Motivated by theoretical results pertaining to combining rankings from different knowledge sources (Colombo et al., 2022; Dwork et al., 2001), we established an overall ranking by simply averaging the per-task ranks.

This approach has several disadvantages. First, it is difficult to incorporate new metrics into the comparison, since this requires not only computing the score of a new metric on 201 tasks, but also comparing it to all existing metrics on each task using expensive resampling significance tests. Adding a new metric also has the undesirable effect of potentially causing other metrics to swap places in the overall rankings. While rank averaging has theoretical underpinnings, as noted above, these apply to settings in which the constituent tasks provide only ranking information themselves. In order to take advantage of richer information available from correlation statistics, we derived dense ranks from pairwise significance tests, but this relies on an ad hoc clustering algorithm, and it is not clear to what extent our average ranks are supported by the original theory. They also lack confidence information, making it difficult to quantify conclusions about the overall superiority of one metric over another.

This year we adopted a much simpler approach in order to address these difficulties. We use just 10 main tasks, and compute an overall score by taking a weighted average of results from each task. We perform significance tests on each pair of metrics for each task as before, but also do so for each pair of metrics on the overall average score, allowing us to establish a clearer global ranking. The average score for a new metric can be computed relatively quickly, and it does not affect the scores of other metrics. Significance tests still require the expensive step of comparing to all other metrics, but they are no longer necessary for computing a metric's raw overall score.

We acknowledge that this approach is not perfect. One problem is that we need to combine correlations and accuracies that may have different dynamic ranges. For example, the mean Pearson correlation across all metrics for en→de at the system level is 0.88 with standard deviation 0.24, while at the segment level it is 0.39 with a standard deviation of 0.17. Averaging system-level and segment-level correlations will therefore effectively upweight the system-level contribution. We experimented with different weightings to compensate for this, but found that they did not make a large differ-

| language | ref used | scored ref |
|---|---|---|
| en→de | A | – |
| he→en | B | A |
| zh→en | A | – |

Table 6: Use of reference translations.

ence, and decided to use equal weights for simplicity. Another problem is that we do not account for dependencies among tasks. Although all tasks are at least somewhat complementary, many–such as system-level and segment-level correlations—are based on the same underlying data, and thus violate the assumptions of our hypothesis tests. We leave more sophisticated inference approaches such as proposed by Dror et al. (2017) or Hagmann and Riezler (2023) for future work.

### 5.1 Task Attributes

Tasks are identified by unique value assignments for each of the following attributes: language, level, and correlation statistic. Unlike last year, we no longer have tasks specific to different domains, as domains differ across languages this year. We also drop the "include-human" vs "no-human" distinction, and always score reference translations that are not used by the metrics. As shown in Table 6, Hebrew→English is the only language pair for which such a reference is available. Finally, last year we used three different averaging methods for each correlation statistic at the segment level; this year we choose only one method for each segment-level correlation.

Attributes are as follows:

**Language**

Language pairs include those for which we have MQM ratings—English→German, Hebrew→English, and Chinese→English—plus *all*, which indicates all pairs pooled together.

**Level**

We computed correlations at the *system* level and the *segment* level. For English→German, segments are paragraphs; for the two other language pairs, they are sentences. System-level scores for human ratings and for all metrics that did not supply an explicit system-level score are averages over segment-level scores.

**Correlation/accuracy**

We computed three correlation/accuracy statistics selected to provide complementary information:

| task | lang | level | correlation | wt |
|---|---|---|---|---|
| 1 | all | system | accuracy | 3 |
| 2 | en→de | system | Pearson | 1 |
| 3 | en→de | segment | Pearson | 1 |
| 4 | en→de | segment | $\text{acc}^*_{eq}$ | 1 |
| 5 | he→en | system | Pearson | 1 |
| 6 | he→en | segment | Pearson | 1 |
| 7 | he→en | segment | $\text{acc}^*_{eq}$ | 1 |
| 8 | zh→en | system | Pearson | 1 |
| 9 | zh→en | segment | Pearson | 1 |
| 10 | zh→en | segment | $\text{acc}^*_{eq}$ | 1 |

Table 7: Tasks and weighting.

- System-level pairwise ranking *accuracy* (as proposed by Kocmi et al., 2021). This is computed over data pooled across all three language pairs.
- Segment-level pairwise ranking accuracy with tie calibration (as proposed by Deutsch et al., 2023). We use the $\text{acc}^*_{eq}$ variant to compare vectors of metric and gold scores for each segment, then average the results over segments.
- System- and segment-level Pearson correlation. At the segment level, we flatten matrices of system × segment scores into vectors before comparing them.

### 5.2 Tasks and Weighting

Table 7 shows the complete list of tasks and their weights. All tasks receive a weight of 1, except for system-level accuracy, which has a weight of 3 because it combines data from all three language pairs.

To compute a global score for each metric across all tasks, we first map Pearson correlations from their natural range of $[-1, 1]$ into the $[0, 1]$ range of the accuracy scores, then take a weighted average of the results.

### 5.3 Rank Assignment

For each task, we assign ranks to metrics based on their significance clusters. To do so, we compare all pairs of metrics and determine whether the difference in their correlation scores is significant, according to the PERM-BOTH hypothesis test of Deutsch et al. (2021). We use 1000 re-sampling runs and set $p = 0.05$. As advocated by Wei et al.

(2022), we divide the sample into blocks of 100, compute significance after each block (cumulative over all blocks sampled so far), and stop early if the p-value is $< 0.02$ or $> 0.50$.

The $\text{acc}^*_{eq}$ statistic creates a problem for significance testing because it optimizes a latent tie threshold for each metric on each test set (just one threshold for all item-wise score vectors). Since the permutation test for comparing two metrics creates two new vectors by randomly swapping elements of the original vectors on each draw, this necessitates the very expensive step of finding two new tie thresholds for each draw. To reduce the expense, we used the following approximate procedure. First find an optimal threshold for each input metric on the current test set, then create all pairs of item-wise scores and assign a correct/incorrect status to each pair by examining whether the metric's ranking matches the human ranking. Then perform the permutation test on these pairwise status vectors rather than the original score vectors. This approximation has more degrees of freedom than the original test, and can sample pairs that would never result from swapping the original score vectors, but our experiments showed that it is a reasonable proxy for the correct procedure.

To compute overall p-values based on weighted average scores of two metrics across all tasks, we cache the results of the draws for the per-task significance tests. In all cases, these are vectors of $K$ pairs of correlation or accuracy statistics. Where $K < 1000$ due to early stopping, we duplicate elements to get 1000 examples. Then for $i$ in 1..1000 we compare the weighted average of the pairs from the $i$th draw across all tasks, and record the results to produce an overall p-value.

### Clustering

Given significance results (p-values) for all pairs of metrics, we assign ranks as follows. Starting with the highest-scoring metric, we move down the list of metrics in descending order by score, and assign rank 1 to all metrics until we encounter the first metric that is significantly different from any that have been visited so far. That metric is assigned rank 2, and the process is repeated. This continues until all metrics have been assigned a rank. Note that this is a greedy algorithm, and hence it can place two metrics that are statistically indistinguishable in different clusters.

## 6 Main Results

As we have seen in Section 5, the main results are the overall scores by taking a weighted average of the results from the ten main tasks, including system-level and segment-level tasks in different translation directions. Similar to last year, since the main use case of automatic metrics is to rank systems, system-level accuracy has a 1/4 weight on the final score with the remaining 3/4 distributed over 9 different settings.

Table 1 shows the official scores and rankings of all baselines and primary submissions. Table 8 and 9 show the scores and rankings of each individual task at system level and segment level, respectively. Similar to last year's results, neural metrics perform significantly better than lexical metrics. Of the 32 evaluated metrics, BLEU, F200SPBLEU and CHRF are ranked 28th, 24th and 29th respectively. On the other hand, fine-tuned neural baseline metrics, like COMET and BLEURT-20, remain ranked higher than several of the new primary submissions. They are surpassed only by submissions relying on significantly larger models.

It is worth noting that the best-performing baseline, COMETKIWI, along with four of the seven top-performing primary submissions, are reference-free. As we will elaborate on in a later section (Section 8), there are quality issues with human reference translations. This highlights the challenge of ensuring robustness to poor-quality references for reference-based metrics. In cases where a high-quality human reference is not available, reference-free metrics can serve as more robust alternatives.

Overall, XCOMET-Ensemble is the best performing metric in terms of average scores over the 10 meta-evaluation settings, with a statistically significant advantage over all other metrics. It consistently correlates best with human MQM scores at segment level for all translation directions, and it is ranked at worst in the 2nd significance cluster for all system-level meta-evaluation tasks.

Figure 1 shows the correlation scores split by translation direction. There are two key observations: 1) a majority of the metrics have higher correlations for en→de among the three translation directions, except for MRE-SCORE-LABSE-REGULAR and EBLEU, that perform substantially better for he→en, and YISI-1 and BERTSCORE, that perform equally in en→de and he→en; 2) reference-based metrics struggle for zh→en due to the reference quality, except for XCOMET-

| Metric | avg-corr | | en→de,he→en,zh→en accuracy task1 | | en→de pearson task2 | | he→en pearson task5 | | zh→en pearson task8 | |
|---|---|---|---|---|---|---|---|---|---|---|
| XCOMET-Ensemble | **1** | **0.825** | **1** | **0.928** | 2 | 0.980 | **1** | **0.950** | 2 | 0.927 |
| XCOMET-QE-Ensemble* | 2 | 0.808 | **1** | **0.908** | 2 | 0.974 | 2 | 0.909 | 3 | 0.892 |
| MetricX-23 | 2 | 0.808 | **1** | **0.908** | 2 | 0.977 | 2 | 0.910 | 4 | 0.873 |
| GEMBA-MQM* | 2 | 0.802 | **1** | **0.944** | **1** | **0.993** | 2 | 0.939 | **1** | **0.991** |
| MetricX-23-QE* | 2 | 0.800 | 2 | 0.892 | 2 | 0.969 | 3 | 0.858 | 4 | 0.859 |
| mbr-metricx-qe* | 3 | 0.788 | 2 | 0.880 | 2 | 0.976 | 2 | 0.915 | 2 | 0.936 |
| MaTESe | 3 | 0.782 | 2 | 0.904 | 4 | 0.918 | 2 | 0.906 | 3 | 0.889 |
| CometKiwi* | 3 | 0.782 | **1** | **0.904** | 3 | 0.946 | 3 | 0.860 | 2 | 0.963 |
| COMET | 3 | 0.779 | 2 | 0.900 | **1** | **0.990** | 2 | 0.940 | 3 | 0.898 |
| BLEURT-20 | 3 | 0.776 | 2 | 0.892 | **1** | **0.990** | 2 | 0.937 | 4 | 0.880 |
| KG-BERTScore* | 3 | 0.774 | 2 | 0.884 | 4 | 0.926 | 2 | 0.908 | 2 | 0.962 |
| sescoreX | 3 | 0.772 | 2 | 0.892 | 3 | 0.952 | 3 | 0.901 | 5 | 0.797 |
| cometoid22-wmt22* | 4 | 0.772 | 2 | 0.880 | 2 | 0.973 | 4 | 0.839 | 2 | 0.940 |
| docWMT22CometDA | 4 | 0.768 | 2 | 0.904 | **1** | **0.990** | 2 | 0.922 | 3 | 0.907 |
| docWMT22CometKiwiDA* | 4 | 0.767 | 2 | 0.900 | 2 | 0.970 | 2 | 0.906 | 2 | 0.965 |
| Calibri-COMET22 | 4 | 0.767 | **1** | **0.904** | 2 | 0.963 | 2 | 0.930 | 4 | 0.863 |
| Calibri-COMET22-QE* | 4 | 0.755 | 2 | 0.863 | 2 | 0.978 | 4 | 0.778 | 2 | 0.934 |
| YiSi-1 | 4 | 0.754 | 2 | 0.871 | 4 | 0.925 | 2 | 0.917 | 4 | 0.823 |
| MS-COMET-QE-22* | 5 | 0.744 | 2 | 0.871 | 3 | 0.959 | 5 | 0.721 | 3 | 0.901 |
| prismRef | 5 | 0.744 | 2 | 0.851 | 4 | 0.920 | **1** | **0.956** | 6 | 0.762 |
| mre-score-labse-regular | 5 | 0.743 | 2 | 0.888 | 3 | 0.942 | **1** | **0.958** | 3 | 0.903 |
| BERTscore | 5 | 0.742 | 2 | 0.871 | 5 | 0.891 | 3 | 0.895 | 5 | 0.810 |
| XLsim | 6 | 0.719 | 2 | 0.855 | 4 | 0.925 | 3 | 0.887 | 5 | 0.796 |
| f200spBLEU | 7 | 0.704 | 3 | 0.819 | 4 | 0.919 | 4 | 0.805 | 6 | 0.772 |
| MEE4 | 7 | 0.704 | 3 | 0.823 | 5 | 0.861 | 3 | 0.879 | 6 | 0.743 |
| tokengram_F | 7 | 0.703 | 3 | 0.815 | 5 | 0.858 | 3 | 0.878 | 5 | 0.795 |
| embed_llama | 7 | 0.701 | 3 | 0.831 | 5 | 0.861 | 4 | 0.841 | 5 | 0.785 |
| BLEU | 7 | 0.696 | 3 | 0.815 | 4 | 0.917 | 5 | 0.769 | 7 | 0.734 |
| chrF | 7 | 0.694 | 3 | 0.795 | 5 | 0.866 | 4 | 0.776 | 5 | 0.809 |
| eBLEU | 7 | 0.692 | 2 | 0.859 | 4 | 0.918 | 2 | 0.911 | 7 | 0.727 |
| Random-sysname* | 8 | 0.529 | 4 | 0.578 | 6 | 0.357 | 6 | 0.209 | 8 | 0.093 |
| prismSrc* | 9 | 0.455 | 5 | 0.386 | 6 | -0.327 | 6 | -0.017 | 8 | -0.406 |

Table 8: Results on system-level tasks for main language pairs. Rows are sorted by the overall average correlation across all 10 tasks (leftmost column). Starred metrics are reference-free, and underlined metrics are baselines.

ENSEMBLE and SESCOREX. The reason for the significant drop in correlation for he→en is unclear. This drop is observed across almost all metrics, whether they are trained or untrained, reference-free or reference-based, and they exhibit varying degrees of degradation.

We continue to be interested in metrics' ability to generalise across domains. In Figure 2, 3 and 4 we present the performance of each metric across different domains in each translation direction. Most metrics perform well in evaluating translation in the user reviews domain across translation direction, despite lacking annotated data in that domain. Further investigation is required to understand whether this is because the translation quality of MT output is more diverse in the user reviews domain, making it easier for metrics to accurately discriminate.

Figure 5 shows the average correlations of metrics when grouped separately by system-level and segment-level tasks. Many metrics fall into the same significance cluster when evaluated on the system-level, as we only have a limited number of MT systems. Although most of the metrics compute the system-level score by averaging their segment-level scores, we observe that high correlations between automatic metrics and human judgments at the segment level do not necessarily guarantee high correlations at the system level. For example, PRISMSRC is in the middle of the pack and has moderate Pearson's correlation at segment level for en→de. However, it is negatively correlating with human judgements when evaluating the same language pair at system level.

## 7 Understanding metrics' scores beyond correlation

In the past few years, we demonstrated that new metrics correlate better with human judgments than BLEU does. Some new baseline metrics even consistently outperform BLEU for consecutive years across translation directions and domains. How-

| Metric | en→de pearson task3 | | en→de acc-t task4 | | he→en pearson task6 | | he→en acc-t task7 | | zh→en pearson task9 | | zh→en acc-t task10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| XCOMET-Ensemble | **1** | **0.695** | **1** | **0.604** | **1** | **0.556** | **1** | **0.586** | **1** | **0.650** | **1** | **0.543** |
| XCOMET-QE-Ensemble* | 2 | 0.679 | 3 | 0.588 | 3 | 0.498 | 4 | 0.554 | **1** | **0.647** | 3 | 0.533 |
| MetricX-23 | 4 | 0.585 | **1** | **0.603** | **1** | **0.548** | 2 | 0.577 | 2 | 0.625 | 3 | 0.531 |
| GEMBA-MQM* | 6 | 0.502 | 5 | 0.572 | 5 | 0.401 | 3 | 0.564 | 6 | 0.449 | 5 | 0.522 |
| MetricX-23-QE* | 3 | 0.626 | 2 | 0.596 | 2 | 0.520 | 3 | 0.564 | **1** | **0.647** | 4 | 0.527 |
| mbr-metricx-qe* | 4 | 0.571 | 3 | 0.584 | 5 | 0.411 | 4 | 0.553 | 5 | 0.489 | 2 | 0.537 |
| MaTESe | 5 | 0.554 | 9 | 0.528 | 4 | 0.459 | 5 | 0.550 | 4 | 0.511 | 12 | 0.479 |
| CometKiwi* | 7 | 0.475 | 5 | 0.569 | 7 | 0.387 | 6 | 0.544 | 6 | 0.442 | 4 | 0.525 |
| COMET | 8 | 0.432 | 4 | 0.574 | 5 | 0.401 | 8 | 0.532 | 8 | 0.396 | 7 | 0.514 |
| BLEURT-20 | 7 | 0.484 | 5 | 0.572 | 8 | 0.382 | 10 | 0.519 | 9 | 0.378 | 6 | 0.518 |
| KG-BERTScore* | 8 | 0.451 | 7 | 0.556 | 8 | 0.382 | 7 | 0.537 | 7 | 0.430 | 6 | 0.516 |
| sescoreX | 5 | 0.519 | 6 | 0.563 | 7 | 0.385 | 15 | 0.484 | 3 | 0.536 | 9 | 0.499 |
| cometoid22-wmt22* | 8 | 0.441 | 4 | 0.578 | 9 | 0.365 | 11 | 0.515 | 5 | 0.479 | 7 | 0.515 |
| docWMT22CometDA | 10 | 0.394 | 7 | 0.559 | 10 | 0.339 | 13 | 0.497 | 10 | 0.353 | 10 | 0.493 |
| docWMT22CometKiwiDA* | 8 | 0.444 | 8 | 0.547 | 12 | 0.286 | 14 | 0.489 | 8 | 0.387 | 10 | 0.493 |
| Calibri-COMET22 | 9 | 0.413 | 10 | 0.522 | 5 | 0.401 | 11 | 0.515 | 8 | 0.396 | 14 | 0.474 |
| Calibri-COMET22-QE* | 8 | 0.441 | 12 | 0.483 | 6 | 0.395 | 12 | 0.506 | 6 | 0.443 | 10 | 0.491 |
| YiSi-1 | 11 | 0.366 | 8 | 0.542 | 6 | 0.395 | 8 | 0.529 | 11 | 0.290 | 8 | 0.504 |
| MS-COMET-QE-22* | 12 | 0.310 | 8 | 0.546 | 12 | 0.295 | 13 | 0.498 | 9 | 0.367 | 9 | 0.498 |
| prismRef | 6 | 0.516 | 10 | 0.518 | 11 | 0.319 | 9 | 0.528 | 14 | 0.183 | 8 | 0.504 |
| mre-score-labse-regular | 17 | 0.111 | 9 | 0.530 | 8 | 0.378 | 10 | 0.522 | 16 | 0.145 | 12 | 0.481 |
| BERTscore | 12 | 0.325 | 9 | 0.528 | 10 | 0.335 | 11 | 0.515 | 12 | 0.236 | 9 | 0.499 |
| XLsim | 13 | 0.239 | 9 | 0.527 | 14 | 0.233 | 16 | 0.480 | 17 | 0.111 | 15 | 0.464 |
| f200spBLEU | 14 | 0.237 | 9 | 0.526 | 14 | 0.230 | 18 | 0.447 | 18 | 0.108 | 13 | 0.476 |
| MEE4 | 16 | 0.202 | 9 | 0.529 | 13 | 0.256 | 19 | 0.441 | 18 | 0.105 | 12 | 0.480 |
| tokengram_F | 15 | 0.227 | 10 | 0.520 | 14 | 0.226 | 17 | 0.461 | 20 | 0.060 | 11 | 0.485 |
| embed_llama | 13 | 0.250 | 12 | 0.483 | 15 | 0.215 | 20 | 0.430 | 15 | 0.161 | 16 | 0.447 |
| BLEU | 16 | 0.192 | 10 | 0.520 | 15 | 0.220 | 19 | 0.442 | 17 | 0.119 | 14 | 0.472 |
| chrF | 14 | 0.232 | 10 | 0.519 | 15 | 0.221 | 17 | 0.460 | 19 | 0.063 | 11 | 0.485 |
| eBLEU | 19 | -0.011 | 11 | 0.512 | 16 | 0.131 | 18 | 0.445 | 22 | -0.084 | 14 | 0.473 |
| Random-sysname* | 18 | 0.064 | 14 | 0.409 | 17 | 0.041 | 20 | 0.428 | 21 | 0.018 | 18 | 0.381 |
| prismSrc* | 9 | 0.425 | 13 | 0.426 | 16 | 0.140 | 19 | 0.441 | 13 | 0.223 | 17 | 0.421 |

Table 9: Results on segment-level tasks for main language pairs. Rows are sorted by the overall average correlation across all 10 tasks (leftmost column in Table 8). Starred metrics are reference-free, and underlined metrics are baselines.

ever, the research community is still reluctant to adopt newer and better automatic MT evaluation metrics in practice. One of the reasons is that MT researchers have established some "common beliefs" about the relationship between BLEU and actual translation quality, and similar intuitions about new metrics have yet to crystallize. Thus, this year, we conduct two additional analyses beyond correlation with human to understand the meaning of the score differences that metrics present with respect to the statistical significance of MT system rankings according to human annotations and metric scores. Our results should *NOT* be used as arguments to forego significance tests or appropriate human evaluation. These analyses only support an intuitive sense of metric score meanings to encourage broader adoption of new automatic MT evaluation metrics.

### 7.1 Correspondence to MQM scores significance

First, we follow Lo et al. (2023a) to study the relationship between statistically significant differences in human scores and the magnitude of metric differences. Specifically, we run a one-sided paired t-test with an equal variance assumption for each system pair on segment-level MQM scores. After that, we fit the corresponding metric score differences and the p-values of the t-test on the MQM scores to an isotonic regression (Robertson et al., 1988), that predicts whether the human MQM score difference will be significant given the metric's score difference. Isotonic regression produces a non-decreasing function where the classifier output can be interpreted as a confidence level.[9] We set $p_{mqm} < 0.05$ as the significance level of MQM

[9] https://scikit-learn.org/stable/modules/isotonic.html

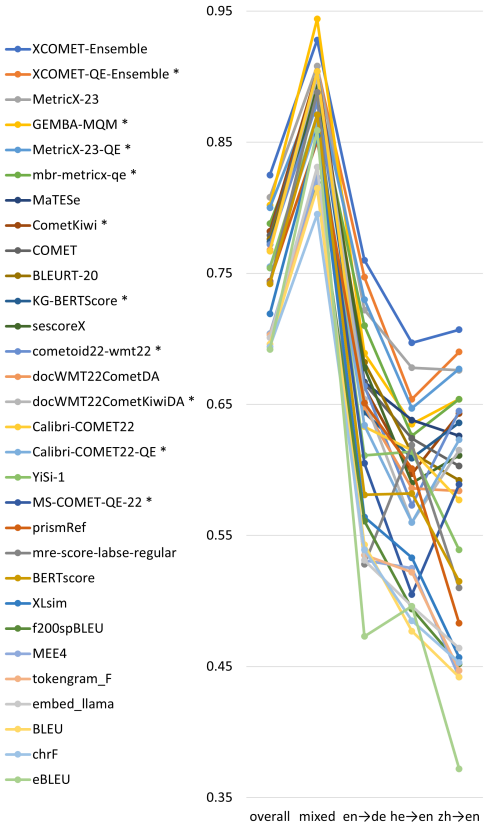


Figure 1: Average metrics' meta-evaluation scores in tasks grouped by translation direction. The "mixed" group is the accuracy score of the metrics in task 1.

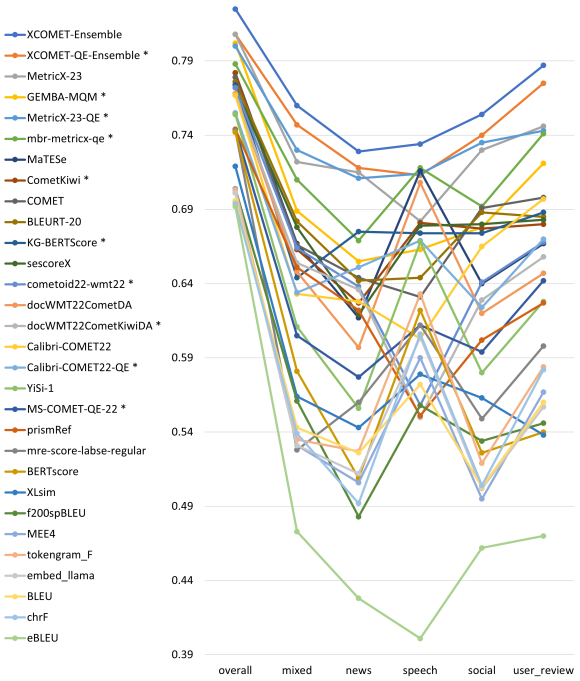


Figure 2: Average metrics' correlation with human in tasks grouped by domain in en→de. The "mixed" group is the average correlation in all en→de tasks.

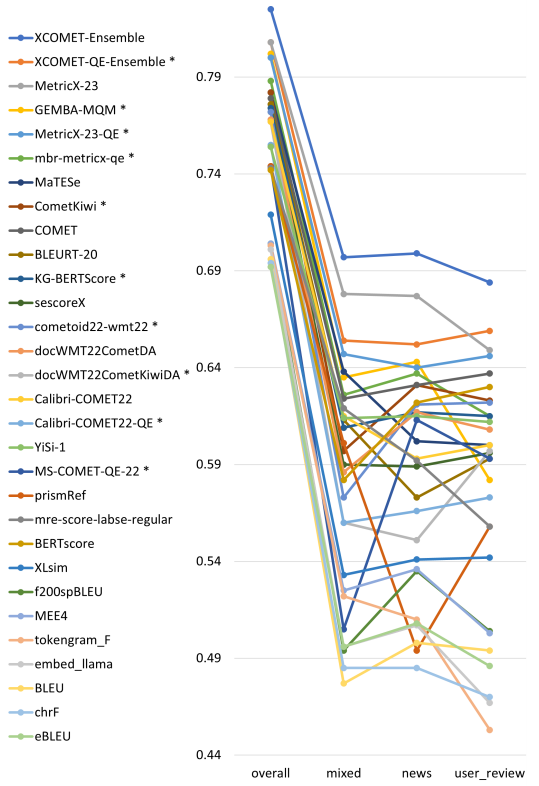


Figure 3: Average metrics' correlation with human in tasks grouped by domain in he→en. The "mixed" group is the average correlation in all he→en tasks.

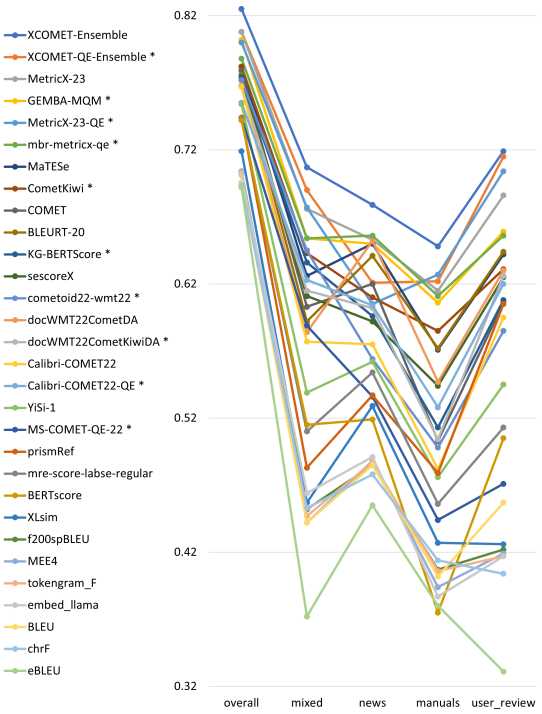


Figure 4: Average metrics' correlation with human in tasks grouped by domain in zh→en. The "mixed" group is the average correlation in all zh→en tasks.

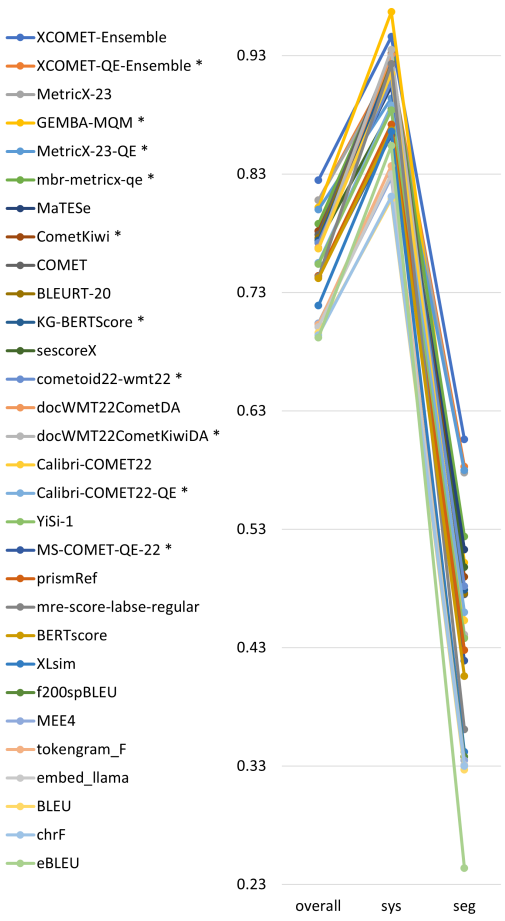


Figure 5: Average metrics' correlation with human in tasks grouped by granularity level.

scores. Thus, the output of the isotonic regression function can be viewed as $Pr(p_{mqm} < 0.05|\Delta M)$ where $p_{mqm}$ is the p-value of the t-test on the MQM scores for each system pair and $\Delta M$ is the metric score difference.

Figure 6 shows the (log) p-value of one-sided paired t-test on the MQM scores against the corresponding BLEU and COMET score difference for each system pair in en→de. Figures 9-14 in appendix D, show the same analyses for all metrics and translation directions. For each metric, we can choose a particular level of confidence (i.e., a point along the y-axis on the right) to give metric score difference cut-offs (i.e., a point along the x-axis) that this metric difference reflects significant MQM score differences. Drawing a horizontal line from the confidence level, say 80%, to the red line enables us to find the minimum metric difference cut-off required at the corresponding x-value down from the red line, i.e. 11 for BLEU in Figure 6. Using this lookup method, Table 10 shows the cut-offs of $\Delta M$ when $Pr(p_{mqm} < 0.05|\Delta M) = 0.8$ for each metric and translation directions.

We run the leave-one-system-out cross validation and Table 10 shows that the range of precision in the cross validation are consistently high across metrics, with the exception of BLEU, CHRF, PRISMSRC, RANDOM-SYSNAME and SLIDE. This means the metric cut-offs we find using the regression model are reliable.

Contrary to the common belief that 2 BLEU improvement represents "significant" or "notable by human" improvement in the actual translation quality, our analyses show that 2.2 BLEU is the minimum required improvement for a high confidence (80%) that MQM annotators to mark significant differences in the translation output for one translation direction (zh→en) and that threshold would be as high as 11 BLEU for en→de. Table 10 serves as a reference between BLEU differences and differences in some of the modern metrics, and assists metric users in understanding scores provided by modern metrics. For example, when evaluating he→en translation quality, we see that a BLEU difference of 3.5 corresponds to 80% confidence that the metric's ranking of the two MT systems will match with the decision made by human annotators with a significant difference. Meanwhile, a COMET score difference of 0.014 would have the same 80% chance of human judged significant difference.

### 7.2 Correspondence to metric scores significance

Inspired by Marie (2022), we run a study similar to that in the previous subsection but on the relations between statistically significant differences in metric scores and the magnitude of metric differences. Instead of one-sided t-test on MQM, the p-values are now obtained by running statistical significance tests with bootstrap resampling on the metric scores for each system pair. Similarly, we fit the corresponding metric score differences and the p-values of the significance test to an isotonic regression for predicting whether the translation quality improvement as indicated by the metric will be significant given the metric score difference. We set $p_M < 0.05$ and thus, the output of the isotonic regression function is now $Pr(p_M < 0.05|\Delta M)$, where $p_M$ is the p-value of the significance test on the metric scores for each system pair and $\Delta M$ is the metric score difference.

Figure 7 shows the (log) p-value of the signifi-

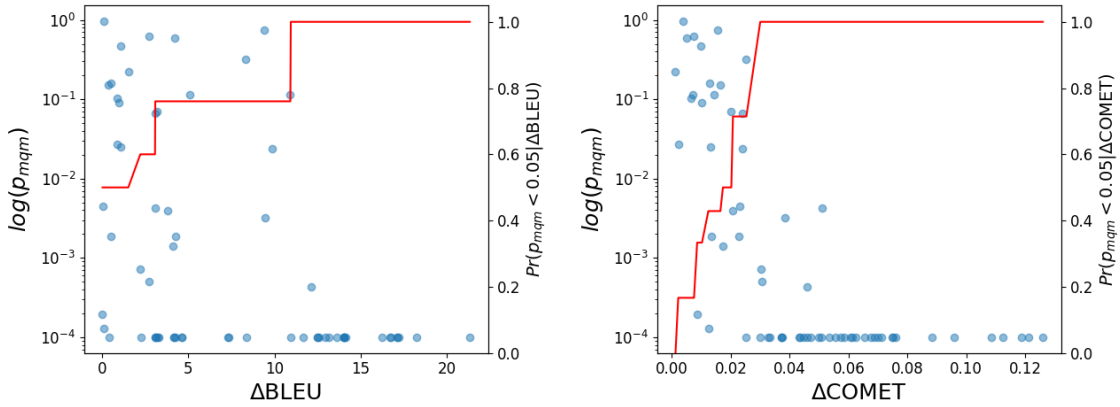


Figure 6: Log p-value of one-sided paired t-test on MQM scores ($p_{mqm}$) against the metric (left: BLEU, right: COMET) score difference for each system pair in en→de. The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of $p_{mqm}$ are rounded up to 0.0001 when they are less than 0.0001.

| | en→de | | he→en | | zh→en | |
|---|---|---|---|---|---|---|
| Metric | min $\Delta M$ | c.v. precision | min $\Delta M$ | c.v. precision | min $\Delta M$ | c.v. precision |
| BERTSCORE | 0.011 | [75-100%] | 0.0053 | [83-100%] | 0.0033 | [75-100%] |
| BLEU | 11 | [33-100%] | 3.5 | [82-100%] | 2.2 | [75-100%] |
| BLEURT-20 | 0.041 | [75-100%] | 0.019 | [100-100%] | 0.013 | [82-100%] |
| CALIBRI-COMET22 | 0.068 | [71-100%] | 0.031 | [89-100%] | 0.043 | [80-100%] |
| CALIBRI-COMET22-QE | 0.072 | [82-100%] | 0.020 | [86-100%] | 0.025 | [67-100%] |
| CHRF | 2.8 | [25-100%] | 3.2 | [83-100%] | 2.6 | [86-100%] |
| COMET | 0.030 | [78-100%] | 0.014 | [88-100%] | 0.013 | [80-100%] |
| COMETKIWI | 0.022 | [67-100%] | 0.014 | [64-100%] | 0.0098 | [62-100%] |
| COMETOID22-WMT22 | 0.018 | [86-100%] | 0.0077 | [71-100%] | 0.011 | [67-100%] |
| DOCWMT22COMETDA | 0.027 | [78-100%] | 0.012 | [82-100%] | 0.014 | [82-100%] |
| DOCWMT22COMETKIWIDA | 0.026 | [75-100%] | 0.012 | [64-100%] | 0.0096 | [71-100%] |
| EBLEU | 0.022 | [57-100%] | 0.019 | [83-100%] | 0.017 | [86-100%] |
| EMBED_LLAMA | 0.062 | [67-100%] | 0.019 | [80-100%] | 0.020 | [80-100%] |
| F200SPBLEU | 4.6 | [60-100%] | 3.6 | [75-100%] | 3.5 | [86-100%] |
| GEMBA-MQM | 2.0 | [89-100%] | 1.0 | [82-100%] | 2.0 | [69-100%] |
| KG-BERTSCORE | 0.0097 | [50-100%] | 0.0097 | [86-100%] | 0.0079 | [62-100%] |
| MATESE | 0.99 | [71-100%] | 0.77 | [75-100%] | 0.70 | [73-100%] |
| MBR-METRICX-QE | 0.047 | [75-100%] | 0.026 | [82-100%] | 0.022 | [75-100%] |
| MEE4 | 0.013 | [71-100%] | 0.024 | [78-100%] | 0.020 | [86-100%] |
| METRICX-23 | 0.73 | [100-100%] | 0.29 | [76-100%] | 0.55 | [83-100%] |
| METRICX-23-QE | 0.53 | [71-100%] | 0.092 | [67-100%] | 0.49 | [60-100%] |
| MRE-SCORE-LABSE-REGULAR | 0.010 | [67-100%] | 0.016 | [100-100%] | 0.0064 | [62-100%] |
| MS-COMET-QE-22 | 1.5 | [80-100%] | 1.4 | [67-100%] | 1.2 | [60-100%] |
| PRISMREF | 0.081 | [75-100%] | 0.14 | [88-100%] | 0.19 | [83-100%] |
| PRISMSRC | 0.036 | [73-100%] | 0.040 | [33-100%] | 0.022 | [64-100%] |
| RANDOM-SYSNAME | 7.8 | [0-100%] | 0.082 | [67-90%] | 5.0 | [50-90%] |
| SESCOREX | 0.38 | [73-100%] | 0.50 | [89-100%] | 0.62 | [73-100%] |
| SLIDE | 0.049 | [78-100%] | 0.017 | [78-100%] | 0.013 | [58-100%] |
| XCOMET-ENSEMBLE | 0.029 | [88-100%] | 0.0092 | [83-100%] | 0.012 | [75-100%] |
| XCOMET-QE-ENSEMBLE | 0.038 | [86-100%] | 0.012 | [83-100%] | 0.021 | [67-100%] |
| XLSIM | 0.015 | [67-100%] | 0.0073 | [82100%] | 0.0091 | [70-100%] |
| YISI-1 | 0.0049 | [67-100%] | 0.0060 | [80-100%] | 0.0054 | [75-100%] |

Table 10: Minimum $\Delta M$ when $Pr(p_{mqm} < 0.05|\Delta M) = 0.8$ for each metric in different translation directions round to 2 significant figures, and the range of precision for the isotonic regression model in leave-one-system-out cross validation.

cance test with bootstrap resampling on the metric scores for BLEU and COMET score difference of each system pair in en→de. Additional figures (Figures 15-20 in appendix Appendix D) show the same analyses for all metrics and translation directions. Using the same lookup method described in the previous subsection, Table 11 shows the cut-offs of $\Delta M$ when $Pr(p_M < 0.05|\Delta M) = 0.8$ for each metric and translation directions.

We run the leave-one-system-out cross valida-

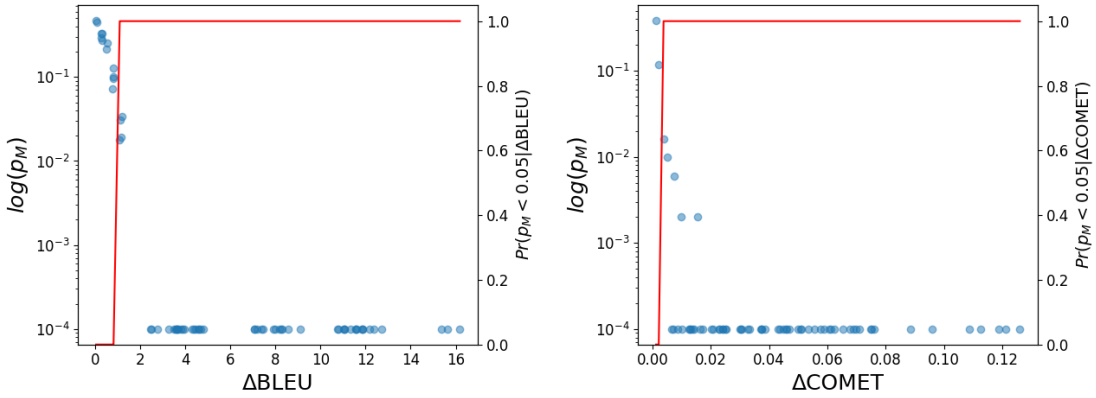


Figure 7: Log p-value of significance test with bootstrap resampling ($p_M$) on system-level metric scores against each metric (left: BLEU, right: COMET) score difference for each system pair in en→de. The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of $p_M$ are rounded up to 0.0001 when they are less than 0.0001.

| | en→de | | he→en | | zh→en | |
|---|---|---|---|---|---|---|
| Metric | min $\Delta M$ | c.v. precision | min $\Delta M$ | c.v. precision | min $\Delta M$ | c.v. precision |
| BERTSCORE | 0.0026 | [100-100%] | 0.0012 | [100-100%] | 0.00085 | [100-100%] |
| BLEU | 1.1 | [100-100%] | 0.79 | [100-100%] | 0.58 | [93-100%] |
| BLEURT-20 | 0.0081 | [100-100%] | 0.0041 | [100-100%] | 0.0024 | [100-100%] |
| CALIBRI-COMET22 | 0.010 | [91-100%] | 0.0063 | [100-100%] | 0.0064 | [100-100%] |
| CALIBRI-COMET22-QE | 0.015 | [100-100%] | 0.0086 | [89-100%] | 0.0078 | [92-100%] |
| CHRF | 0.99 | [100-100%] | 0.68 | [100-100%] | 0.48 | [100-100%] |
| COMET | 0.0038 | [100-100%] | 0.0038 | [90-100%] | 0.0029 | [100-100%] |
| COMETKIWI | 0.0074 | [91-100%] | 0.0019 | [100-100%] | 0.0025 | [93-100%] |
| COMETOID22-WMT22 | 0.0062 | [82-100%] | 0.0026 | [100-100%] | 0.0019 | [100-100%] |
| DOCWMT22COMETDA | 0.0033 | [100-100%] | 0.0013 | [100-100%] | 0.0023 | [100-100%] |
| DOCWMT22COMETKIWIDA | 0.0028 | [100-100%] | 0.0021 | [100-100%] | 0.0015 | [100-100%] |
| EBLEU | 0.0076 | [90-100%] | 0.0048 | [100-100%] | 0.0050 | [100-100%] |
| EMBED_LLAMA | 0.013 | [100-100%] | 0.0079 | [100-100%] | 0.0054 | [100-100%] |
| F200SPBLEU | 1.0 | [100-100%] | 0.94 | [100-100%] | 0.65 | [100-100%] |
| GEMBA-MQM | 0.52 | [100-100%] | 0.38 | [100-100%] | 0.35 | [100-100%] |
| KG-BERTSCORE | 0.0051 | [100-100%] | 0.0016 | [100-100%] | 0.00029 | [93-100%] |
| MATESE | 0.33 | [100-100%] | 0.20 | [100-100%] | 0.15 | [100-100%] |
| MBR-METRICX-QE | 0.0073 | [100-100%] | 0.0039 | [100-100%] | 0.0023 | [100-100%] |
| MEE4 | 0.0029 | [90-100%] | 0.0067 | [100-100%] | 0.0054 | [100-100%] |
| METRICX-23 | 0.23 | [100-100%] | 0.083 | [90-100%] | 0.089 | [92-100%] |
| METRICX-23-QE | 0.19 | [100-100%] | 0.072 | [89-100%] | 0.11 | [100-100%] |
| MRE-SCORE-LABSE-REGULAR | 0.0034 | [100-100%] | 0.0028 | [100-100%] | 0.0010 | [100-100%] |
| MS-COMET-QE-22 | 0.49 | [100-100%] | 0.45 | [88-100%] | 0.18 | [100-100%] |
| PRISMREF | 0.018 | [100-100%] | 0.031 | [100-100%] | 0.020 | [100-100%] |
| PRISMSRC | 0.028 | [100-100%] | 0.025 | [75-100%] | 0.016 | [100-100%] |
| RANDOM-SYSNAME | 0.21 | [100-100%] | 0.14 | [100-100%] | 0.12 | [100-100%] |
| SESCOREX | 0.039 | [100-100%] | 0.10 | [100-100%] | 0.085 | [100-100%] |
| XCOMET-ENSEMBLE | 0.010 | [90-100%] | 0.0035 | [100-100%] | 0.0033 | [100-100%] |
| XCOMET-QE-ENSEMBLE | 0.0065 | [100-100%] | 0.0027 | [100-100%] | 0.0042 | [93-100%] |
| XLSIM | 0.0019 | [100-100%] | 0.0018 | [100-100%] | 0.0022 | [100-100%] |
| YISI-1 | 0.0013 | [100-100%] | 0.0033 | [73-100%] | 0.00074 | [100-100%] |

Table 11: Minimum $\Delta M$ when $Pr(p_M < 0.05|\Delta M) = 0.8$ for each metric in different translation directions round to 2 significant figures, and the range of precision for the isotonic regression model in leave-one-system-out cross validation.

tion, and Table 11 shows that the range of precision in the cross validation are consistently high across metrics. This means the metric cut-offs we find using the regression model are reliable.

Our results, agreeing with Marie (2022), show that to claim significant differences ($p_M < 0.05$) in BLEU with high confidence (80%), the BLEU differences should be greater than 1.1 BLEU for en→de. Table 11 serves as a reference of metric differences with respect to statistical significance with high confidence. For example, when evaluating en→de translation quality, we see that a BLEU

difference of 1.1 corresponds to 80% confidence the difference is statistical significant. Meanwhile, a COMET score difference of 0.0038 would have the same 80% chance of statistical significance.

We have to emphasize again that our result should *NOT* be interpreted as evidence to forego significance tests or appropriate human evaluation. Instead, we are only providing assistance to build an intuition on the meaning of the scores provided by the new metrics to encourage the transition away from BLEU.

## 8 Synthetic Reference Translations

Reference-based metrics compare machine translations of source segments to human translations of those same source segments to determine how good they are. The quality of the underlying human translation is crucial and can impact the quality of the predicted scores more than the choice of metric (Freitag et al., 2020). Motivated by the low human ratings of refA for Chinese→English (Table 4) and the relatively high rankings of reference-free metrics (in comparison to other language pairs) for this language-pair, we investigate a method for generating a synthetic reference translation based on the MT output and the corresponding MQM ratings.

### 8.1 Synthetic Reference Generation

The main idea is straightforward: Given the set of translations of WMT23 General MT Shared Task (generalMT2023) from the WMT campaign and their corresponding MQM ratings, we generate a new synthetic reference translation by choosing for each segment the translation that received the lowest MQM error score as the selected reference. The original human reference translation (i.e. refA) is considered as one of the possible translations in this process, and MQM score ties are broken randomly. Table 12 shows the resulting MQM score of the synthetic reference translations. We were able to reduce the MQM score to below 1 for both tested language pairs (en→de and zh→en), which corresponds to an average of less than one minor error per segment. While this may seem like a significant improvement, we must caution the reader that this is in essence "cherry-picking" based on the MQM ratings and may therefore introduce many hidden issues.

It is also interesting to understand how many segments come from each of the individual MT systems in this selection process. Table 13 shows the number of segments contributed by each system to the generated synthetic reference translations. Unsurprisingly, the top performing MT systems are also the main contributors to the selected synthetic reference translation. For en→de, refA (the original human-generated reference translation) provided the majority of the selected translations, while for zh→en GPT4-5shot is the main contributor, reflecting that the human-generated reference refA for zh→en was indeed error-prone. However, it is interesting to note that despite the overall low quality of this human-generated reference, our method still selected 209 segments from this translation as the lowest-error translation. This would appear to indicate that these human-generated reference translations are not uniformly bad, and only a subset of the translations were unreliable and contained major errors. A possible explanation could be that multiple translators worked on the reference, however, we confirmed with the sponsor translating zh→en that all segments were translated with the same translator.

| | zh→en | en→de |
|---|---|---|
| synthetic Ref. | 0.66 | 0.87 |
| best MT | 2.10 | 3.72 |
| refA | 4.83 | 2.96 |

Table 12: MQM scores of the synthetic references.

| zh→en | | en→de | |
|---|---|---|---|
| GPT4-5shot | 314 | refA | 243 |
| refA | 209 | GPT4-5shot | 57 |
| Lan-BridgeMT | 157 | ONLINE-B | 36 |
| ANVITA | 142 | ONLINE-A | 20 |
| HW-TSC | 105 | AIRC | 20 |
| IOL_Research | 42 | ONLINE-W | 19 |
| ONLINE-W | 33 | NLLB_Greedy | 14 |
| ONLINE-Y | 28 | NLLB_MBR_BLEU | 13 |
| ONLINE-B | 26 | ONLINE-G | 10 |
| ONLINE-A | 24 | ONLINE-Y | 9 |
| ONLINE-G | 21 | Lan-BridgeMT | 9 |
| NLLB_Greedy | 20 | ONLINE-M | 8 |
| ZengHuiMT | 18 | ZengHuiMT | 2 |
| Yishu | 18 | | |
| NLLB_MBR_BLEU | 14 | | |
| ONLINE-M | 6 | | |

Table 13: Number of segments contributed by each system towards the synthetic reference.

### 8.2 Impact on Metrics

Figure 8 compares the segment-level and system-level Pearson correlations of all submitted metrics

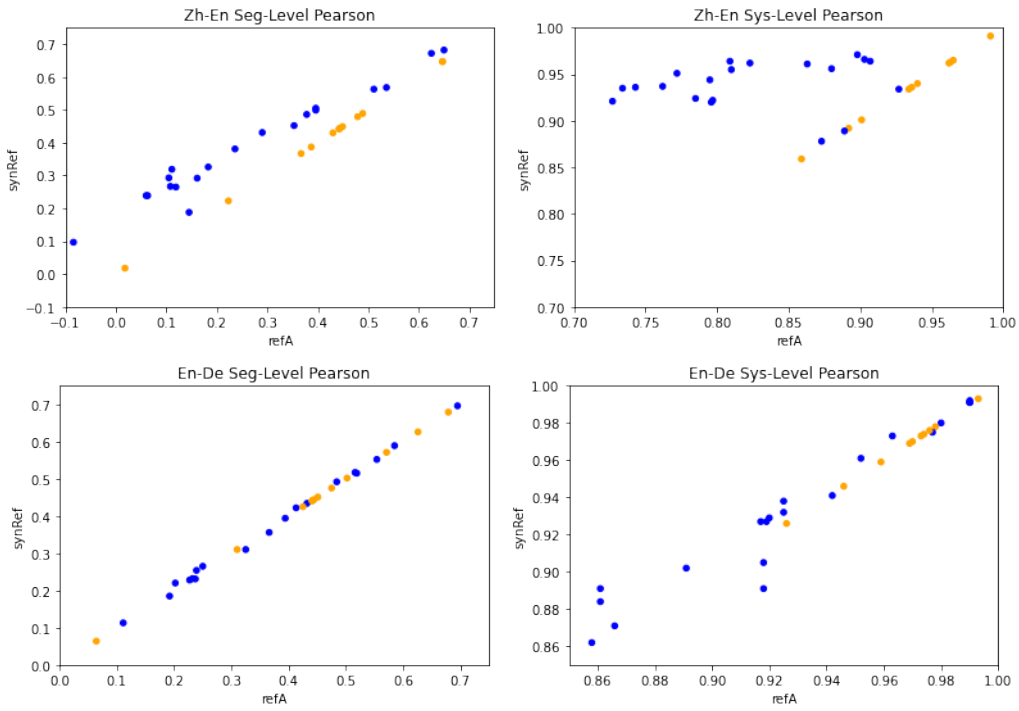


Figure 8: Pearson Correlation when using either the synthetic ref or the original human translation as reference translation. QE metrics are coloured in yellow; ref-based metrics are coloured in blue.

when using either the original or the synthetic reference translation. Reference-based metrics are coloured in blue, while QE metrics are coloured in orange. Obviously, as QE metrics do not use reference translations, their correlations are exactly the same. For Chinese→English, replacing the human-generated reference translation by the synthetic reference translation has a dramatic impact. All reference-based metrics increased their correlation levels with human judgements at both the segment-level and the system-level. This clearly indicates how critically important a high quality reference translation is for reference-based metrics, but moreover, it also highlights the advantages of QE metrics in cases where human-generated references have major quality issues. For the English→German language pair, the human-generated reference translation is of higher quality than any submitted MT system. Consequently, the synthetic reference translation had almost no impact on the segment-level correlations and only a mixed impact on the system-level correlations.

The main takeaways from this study are (i) poor human-generated reference translations can dramatically hurt the performance and reliability of your metric, (ii) strong QE metrics can be better alternatives in such scenarios, and (iii) generating a synthetic reference translation from all system outputs can be used to mitigate bad reference translations, although it assumes obtaining MQM annotations and suffers from cherry-picking bias.[10]

An open unanswered question remains: is it always necessary for a reference translation to be of higher quality than the translation generated by the MT system, in order to have a reliable reference-based metric? This would imply that generating a synthetic reference translation with any errors is problematic, since for any reference-based automatic metric, these synthetic references would become useless for evaluating any MT system that generates translations that surpasses the reference in quality.

## 9 DA+SQM Human Evaluation

In addition to our MQM annotations and as a contrastive evaluation to cover more language pairs, we look into the performance of metrics when compared to the human evaluation campaign conducted by the WMT23 General MT Shared Task (Kocmi et al., 2023), who ran human evaluation for all 14 translation directions and all WMT23 submissions.

In contrast to previous years, they no longer use

[10]Among other issues, any practical strategy for creating synthetic references would need to have a way of avoiding bias toward systems that are similar to the ones used for reference creation.

MTurk neither reference-based evaluation for into-English language pairs. They no longer use z-score normalization because the user interface decision to not track users (i.e., only maintaining HIT information) means that the z-scores are likely to be influenced by the distribution of system quality in the HITs rather than only annotator variation.

They employ the Direct Assessment Scalar Quality Metrics (DA+SQM) technique as presented in Kocmi et al. (2022a).

**DA+SQM** asks bilingual raters to annotate system translations against original sources on a 0–100 labelled scale. The scale is marked with seven points representing expected quality.

At the time of writing, the WMT23 General MT Shared Task had collected data only for 8 translation directions: Chinese↔English (zh↔en), German↔English (de↔en), Japanese↔English (ja↔en), English→Czech (en→cz), and Czech→Ukrainian (cz→uk).

We present system-level accuracy results for both MQM and DA+SQM in Table 14. There are many factors that could affect the ranking. Apart from using a different human annotation protocol, MQM compares 3 translation directions whereas the DA+SQM compares 8 translation directions, containing also the non-English low-resource pair of cz→uk. There is an overlap of only two translation directions between the two: en→de and zh→en. The main difference in ranking is for metrics XCOMET-Ensemble and MetricX-23 ranking significantly lower than for MQM. Investigating system-level Pearson's correlation for individual languages in Tables 19 to 27 shows that both metrics are performing considerably lower across all languages (except en→cz and cz→uk) and we do not see any pattern behind the drop in performance.

| Metric | MQM | | DA+SQM | |
|---|---|---|---|---|
| Translation directions | 3 | | 8 | |
| System pairs (N) | 237 | | 793 | |
| GEMBA-MQM* | 0.944 | (1) | 0.899 | (1) |
| XCOMET-Ensemble | 0.928 | (2) | 0.870 | (10) |
| MetricX-23 | 0.908 | (3) | 0.863 | (11) |
| XCOMET-QE-Ensemble* | 0.908 | (4) | 0.871 | (8) |
| CometKiwi* | 0.904 | (5) | 0.887 | (3) |
| COMET | 0.900 | (6) | 0.890 | (2) |
| BLEURT-20 | 0.892 | (7) | 0.880 | (6) |
| MetricX-23-QE* | 0.892 | (8) | 0.870 | (9) |
| mre-score-labse-regular | 0.888 | (9) | 0.861 | (12) |
| KG-BERTScore* | 0.884 | (10) | 0.884 | (4) |
| cometoid22-wmt22* | 0.880 | (11) | 0.884 | (5) |
| BERTscore | 0.871 | (12) | 0.799 | (16) |
| MS-COMET-QE-22* | 0.871 | (13) | 0.879 | (7) |
| YiSi-1 | 0.871 | (14) | 0.832 | (13) |
| eBLEU | 0.859 | (15) | 0.781 | (19) |
| XLsim | 0.855 | (16) | 0.831 | (14) |
| prismRef | 0.851 | (17) | 0.808 | (15) |
| embed_llama | 0.831 | (18) | 0.778 | (20) |
| f200spBLEU | 0.819 | (19) | 0.786 | (17) |
| BLEU | 0.815 | (20) | 0.770 | (22) |
| tokengram_F | 0.815 | (21) | 0.786 | (18) |
| chrF | 0.795 | (22) | 0.777 | (21) |
| Random-sysname* | 0.578 | (23) | 0.580 | (23) |
| prismSrc* | 0.386 | (24) | 0.412 | (24) |

Table 14: Comparison between system-level pairwise accuracy using MQM and DA+SQM gold scores. MQM results pool data from our 3 main language pairs; DA+SQM results pool data from the 8 language pairs for which DA+SQM scores are available. Rows are sorted by MQM accuracy, with the pure rank order indicated in brackets. Starred metrics are reference-free and underlined metrics are baselines.

## 10 Challenge Sets Sub-task

For the second year, we included a sub-task on challenge sets. This sub-task is inspired by the *Build it or break it: The Language Edition* shared task (Ettinger et al., 2017) which aimed at testing the generalizability of NLP systems beyond the distributions of their training data. Whereas the standard evaluation of the shared task runs on test sets containing generic text from real-world content, the challenge set evaluation is based on test sets designed with the aim of revealing the abilities or the weaknesses of the metrics on evaluating particular translation phenomena. In order to shed light on different perspectives on evaluation, the sub-task takes place in a decentralized manner, where, contrary to the main metric task, the test sets are not provided by the organizers but by different research teams, who are also responsible for analysing and presenting the results.

This subtask is made of three consecutive phases; 1) the *Breaking Round*, 2) the *Scoring Round* and 3) the *Analysis Round*:

1. In the *Breaking Round*, every challenge set participant (*Breaker*) submits their challenge set $S$ composed of contrastive examples for different phenomena, where every example $(s, \hat{t}, t, r) \in S$ contains one source sentence $s$, one incorrect translation $\hat{t}$, one correct translation $t$ and one reference $r$.

2. In the *Scoring Round*, the organizers decompose the $S$ into a blind test set $S'$, where each example includes either an incorrect translation $(s, \hat{t}, r)$ or a correct translation $(s, t, r)$ along with the source and the reference. The separated contrastive examples are shuffled, and the golden truth of which samples are correct or incorrect is kept in a separate set. The

| challenge set | directions | phenomena | items | citation | availability (https://github.com/) |
|---|---|---|---|---|---|
| ACES | 146 | translation errors | 36476 | Amrhein et al. (2023) | EdinburghNLP/ACES |
| DFKI-CS | 3 | linguistic phenomena | 20993 | Avramidis et al. (2023) | DFKI-NLP/mt-testsuite |
| MSLC23 | 4 | low quality MT | 9345 | Lo et al. (2023b) | nrc-cnrc/MSLC23 |

Table 15: Overview of the participation at the metrics challenge sets sub-task

metrics participants from the main task (the *Builders*) are asked to score with their metrics the translations in the given blind test set without knowing which ones are correct or incorrect. Also, in this phase, the organizers score all data with the baseline metrics.

3. Finally, after having gathered all metric scores, the organizers return the respective scored translations to the *Breakers* for the *Analysis round*, where they look at which metrics are able to correctly rank the correct translations higher than the incorrect ones for the phenomena being tested.

There were 3 submissions this year, covering a wide range of phenomena and 146 different translation directions. An overview of the submitted challenge sets can be seen in Table 15. A short description of every submission follows:

**ACES Challenge Set** The Translation Accuracy ChallengE Set (ACES, Amrhein et al., 2023) consists of 36K examples representing challenges from 68 phenomena and covering 146 translation directions. The phenomena range from simple perturbations at the word/character level to more complex errors based on discourse and real-world knowledge. We benchmark the performance of segment-level metrics submitted to WMT 2023 using ACES. For each metric, the authors provide a detailed profile of performance across the ten top-level accuracy error categories in ACES as well as an overall *ACES-Score* for quick comparison. They also measure the incremental performance of the metrics submitted to both WMT 2023 and 2022.

They find that:

- there is no clear *winner* among the metrics submitted to WMT 2023,
- neural metrics also tend to focus more on lexical overlap than semantic content,
- reference-free metrics using language-agnostic multilingual embeddings struggle with detecting untranslated or sentences translated in the wrong direction, and
- performance change between the 2023 and 2022 versions of the metrics is highly variable.

The authors' recommendations are similar to those from WMT 2022. Metric developers should focus on: building ensembles of metrics from different design families, developing metrics that pay more attention to the source and rely less on surface-level overlap, and carefully determining the influence of multilingual embeddings on MT evaluation.

**DFKI Challenge Set** The submission by DFKI (Avramidis et al., 2023) employs a linguistically motivated challenge set that includes about 21,000 items extracted from 155 machine translation systems for three language directions (de→en, en→de, en→ru), covering more than 100 linguistically-motivated phenomena organized in 14 categories. The metrics that have the best performance with regard to our linguistically motivated analysis are the COMETOID22-WMT23 for de→en and METRICX-23-C for en→de and en→ru. Some of the most difficult phenomena for the metrics to score are *passive voice* for de→en, *named entities, terminology* and *measurement units* for en→de and *focus particles, adverbial clause* and *stripping* for en→ru.

**MSLC23 Challenge Set** The Metric Score Landscape Challenge (MSLC23; Lo et al., 2023b) data set aims to gain insight into metric scores on a broader/wider landscape of MT quality. Recent development of MT evaluation metrics has focused on improving their correlation with human judgment on translations of high-quality systems (e.g., participants in the WMT News/General MT Shared Tasks). This means that metric performance may be untested on low- to medium-quality MT output. MSLC23 provides a collection of low- to medium-quality MT output on the news portion of the WMT23 General MT Shared Task test set. Together with the high quality systems submitted to the General MT Shared Task, this enables better interpretation of metric scores across a range of different levels of translation quality. With this

wider range of MT quality, the authors also visualize and analyse metric characteristics beyond just correlation.

The authors find that the smaller variations in segment-level scores given by some metrics at the low end of quality could indicate that these metrics struggle to discriminate between low-quality MT systems. This is further shown by the observation that some metrics rank the low-quality systems in reverse order at system level. A "universal score" phenomenon for some metrics, where a small subset of non-minimum/maximum distinct scores are assigned to a variety of translation output, has been discovered. There is also an observation of diverse behaviours from different metrics on empty string translation. These results highlight the need for metric researchers to check their metrics' performance on a wider landscape of translation quality, or to indicate to potential users that they should be cautious about using their metric on a wide range of quality.

## 11 Conclusion

This paper summarizes the results of the WMT23 shared task on automated machine translation evaluation, the Metrics Shared Task. We presented an extensive analysis on how well metrics perform on our three main translation directions: English→German, Hebrew→English and Chinese→English. The results, based on 10 different tasks, confirm the superiority of neural-based learned metrics over overlap-based metrics like BLEU, SPBLEU or CHRF. These results are confirmed with DA+SQM human judgement. We also found that reference-free metrics were strong contenders this year, partly because they do not rely on the quality of reference translations, an increasingly important issue as MT systems under evaluation become better. In addition, we continued the challenge set subtask, where participants had to create contrastive test suites for evaluating metrics' ability to capture and penalise specific types of translation errors.

## 12 Ethical Considerations

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

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 this year's shared task and publicly available. Newer versions of COMET were developed without using any of the test set, test suite or challenge sets. We ensured that the metrics co-authored by Tom Kocmi were implemented without using any privileged test sets or insider information.

## 13 Acknowledgments

Results for this shared task would not be possible without tight collaboration with the organizers of the WMT23 General MT Shared Task. We are grateful to Google and Unbabel for sponsoring and overseeing the human evaluation.

Ricardo Rei is supported by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI.

Eleftherios Avramidis is supported by the German Research Foundation (DFG) through the project TextQ (grant num. MO 1038/31-1, 436813723), and by the German Federal Ministry of Education and Research (BMBF) through the project SocialWear (grant num. 01IW2000).

## References

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2023. ACES: Translation accuracy challenge sets at wmt 2023. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu,

Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report. *arXiv preprint arXiv:2305.10403*.

Eleftherios Avramidis, Shushen Manakhimova, Vivien Macketanz, and Sebastian Möller. 2023. Challenging the state-of-the-art machine translation metrics from a linguistic perspective. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

Frédéric Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orăsan, and André F. T. Martins. 2023. Findings of the WMT 2023 Shared Task on Quality Estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stéphan Clémençon. 2022. What are the best systems? New perspectives on NLP Benchmarking. *arXiv preprint arXiv:2202.03799*.

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

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties Matter: Modifying Kendall's Tau for Modern Metric Meta-Evaluation. *arXiv preprint arXiv:2305.14324*.

Sören Dreano, Derek Molloy, and Noel Murphy. 2023a. Embed_Llama: using LLM embeddings for the Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Sören Dreano, Derek Molloy, and Noel Murphy. 2023b. Tokengram_F, a fast and accurate token-based chrF++ derivative. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. 2001. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, page 613–622, New York, NY, USA. Association for Computing Machinery.

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

Bryan Eikema and Wilker Aziz. 2021. Sampling-Based Minimum Bayes Risk Decoding for Neural Machine Translation. *arXiv preprint arXiv:2108.04718*.

Muhammad ElNokrashy and Tom Kocmi. 2023. eBLEU: Unexpectedly Good Machine Translation Evaluation Using Simple Word Embeddings. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond English-Centric Multilingual Machine Translation. *Journal of Machine Learning Research*, 22(1).

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

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

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

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon Sampling Rocks: Investigating Sampling Strategies for Minimum Bayes Risk Decoding for Machine Translation. *arXiv preprint arXiv:2305.09860*.

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.

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

Thamme Gowda, Tom Kocmi, and Marcin Junczys-Dowmunt. 2023. Cometoid: Distilling Strong Reference-based Machine Translation Metrics into Even Stronger Quality Estimation Metrics. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *arXiv preprint arXiv:2310.10482*.

Michael Hagmann and Stefan Riezler. 2023. Towards inferential reproducibility of machine learning research. *arXiv preprint arXiv:2302.04054*.

John Hewitt, Christopher Manning, and Percy Liang. 2022. Truncation sampling as language model desmoothing. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3414–3427, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

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

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

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

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

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022b. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). 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, Rebecca Knowles, and Cyril Goutte. 2023a. Beyond correlation: Making sense of the score differences of new mt evaluation metrics. In *Proceedings of Machine Translation Summit XIX Vol. 1: Research Track*, pages 186–199.

Chi-kiu Lo, Samuel Larkin, and Rebecca Knowles. 2023b. Metric score landscape challenge (MSLC23): Understanding metrics' performance on a wider landscape of translation quality. In *Proceedings of the Eighth Conference on Machine Translation*, Singapore. 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.

Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint arXiv:2303.13809*.

Benjamin Marie. 2022. Yes, we need statistical significance testing. towardsai.net https://pub.towardsai.net/yes-we-need-statistical-significance-testing-927a8d21f9f0.

Ananya Mukherjee and Manish Shrivastava. 2023. MEE4 and XLsim: IIIT HYD's Submissions for WMT23 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum Bayes risk decoding in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 259–272, Online. Association for Computational Linguistics.

Subhajit Naskar, Daniel Deutsch, and Markus Freitag. 2023. Quality Estimation using Minimum Bayes Risk. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

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

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

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

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

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.

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

Vikas Raunak, Tom Kocmi, and Matt Post. 2023. SLIDE: Sliding Document Evaluator for Document-Context Evaluation in Machine Translation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

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

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up COMETKIWI: Unbabel-IST 2023 Submission for the Quality Estimation Shared Task. In *Proceedings of the eighth Conference on Machine Translation*, Singapore. Association for Computational Linguistics.

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

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

T. Robertson, F.T. Wright, and R. Dykstra. 1988. *Order Restricted Statistical Inference*. Probability and Statistics Series. Wiley.

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

Brian Thompson and Matt Post. 2020a. 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.

Brian Thompson and Matt Post. 2020b. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Vasiliy Viskov, George Kokush, Daniil Larionov, Steffen Eger, and Alexander Panchenko. 2023. Semantically-Informed Regressive Encoder Score. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Johnny Wei, Tom Kocmi, and Christian Federmann. 2022. Searching for a higher power in the human evaluation of MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 129–139, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhanglin Wu, Yilun Liu, Min Zhang, Xiaofeng Zhao, Junhao Zhu, Ming Zhu, Xiaosong Qiao, Jingfei Zhang, Ma Miaomiao, Zhao Yanqing, Song Peng, shimin tao, Hao Yang, and Yanfei Jiang. 2023. Empowering a Metric with LLM-assisted Named Entity Annotation: HW-TSC's Submission to the WMT23 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore, Singapore (Hybrid). Association for Computational Linguistics.

Jing Xu, Megan Ung, Mojtaba Komeili, Kushal Arora, Y-Lan Boureau, and Jason Weston. 2023a. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13557–13572, Toronto, Canada. Association for Computational Linguistics.

Wenda Xu, Xian Qian, Mingxuan Wang, Lei Li, and William Yang Wang. 2023b. Sescore2: Learning text generation evaluation via synthesizing realistic mistakes.

Wenda Xu, Yilin Tuan, Yujie Lu, Michael Saxon, Lei Li, and William Yang Wang. 2022. Not all errors are equal: Learning text generation metrics using stratified error synthesis. *arXiv preprint arXiv:2210.05035*.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023c. Instructscore: Explainable text generation evaluation with finegrained feedback.

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 Correlations with MQM for all metrics

Tables 16 and 17 contain system- and segment-level results for all metrics (including contrastive submissions) on the 10 standard tasks described in Table 7. No pairwise significance tests were carried out for these results, so the per-task ranks only indicate each metric's order on that task, rather than its significance cluster as in Tables 8 and 9.

| lang:<br>corr_fcn:<br>metric | avg-corr | | en→de,he→en,zh→en<br>accuracy<br>task1 | | en→de<br>pearson<br>task2 | | he→en<br>pearson<br>task5 | | zh→en<br>pearson<br>task8 | |
|---|---|---|---|---|---|---|---|---|---|---|
| XCOMET-Ensemble | **1** | **0.825** | 6 | 0.928 | 9 | 0.980 | 4 | 0.950 | 14 | 0.927 |
| *XCOMET-XXL* | 2 | 0.824 | 5 | 0.932 | 7 | 0.982 | **1** | **0.964** | 16 | 0.911 |
| *MetricX-23-QE-b** | 3 | 0.823 | 2 | 0.940 | 8 | 0.982 | 5 | 0.947 | 15 | 0.926 |
| *XCOMET-XL* | 4 | 0.816 | 7 | 0.924 | 18 | 0.973 | 11 | 0.937 | 26 | 0.884 |
| *MetricX-23-QE-c** | 5 | 0.813 | 4 | 0.932 | 20 | 0.972 | 8 | 0.939 | 4 | 0.974 |
| *MetricX-23-b* | 6 | 0.811 | 9 | 0.916 | 4 | 0.990 | 15 | 0.928 | 19 | 0.902 |
| XCOMET-QE-Ensemble* | 7 | 0.808 | 13 | 0.908 | 16 | 0.974 | 23 | 0.909 | 23 | 0.892 |
| MetricX-23 | 8 | 0.808 | 12 | 0.908 | 12 | 0.977 | 22 | 0.910 | 28 | 0.873 |
| GEMBA-MQM* | 9 | 0.802 | **1** | **0.944** | **1** | **0.993** | 9 | 0.939 | **1** | **0.991** |
| MetricX-23-QE* | 10 | 0.800 | 24 | 0.892 | 22 | 0.969 | 35 | 0.858 | 30 | 0.859 |
| *cometoid22-wmt23** | 11 | 0.794 | 3 | 0.936 | 10 | 0.979 | 16 | 0.928 | 8 | 0.956 |
| mbr-metricx-qe* | 12 | 0.788 | 29 | 0.880 | 13 | 0.976 | 19 | 0.915 | 11 | 0.936 |
| *CometKiwi-XXL** | 13 | 0.786 | 11 | 0.912 | 6 | 0.986 | 14 | 0.929 | 2 | 0.978 |
| *CometKiwi-XL** | 14 | 0.786 | 8 | 0.916 | 14 | 0.975 | 29 | 0.900 | 3 | 0.974 |
| MaTESe | 15 | 0.782 | 17 | 0.904 | 36 | 0.918 | 25 | 0.906 | 25 | 0.889 |
| CometKiwi* | 16 | 0.782 | 16 | 0.904 | 27 | 0.946 | 34 | 0.860 | 6 | 0.963 |
| COMET | 17 | 0.779 | 20 | 0.900 | 3 | 0.990 | 7 | 0.940 | 21 | 0.898 |
| *MetricX-23-c* | 18 | 0.778 | 10 | 0.916 | 28 | 0.944 | 6 | 0.946 | 9 | 0.953 |
| *instructscore* | 19 | 0.777 | 22 | 0.896 | 25 | 0.952 | 21 | 0.910 | 31 | 0.825 |
| BLEURT-20 | 20 | 0.776 | 23 | 0.892 | 5 | 0.990 | 12 | 0.937 | 27 | 0.880 |
| KG-BERTScore* | 21 | 0.774 | 27 | 0.884 | 30 | 0.926 | 24 | 0.908 | 7 | 0.962 |
| sescoreX | 22 | 0.772 | 25 | 0.892 | 26 | 0.952 | 28 | 0.901 | 35 | 0.797 |
| cometoid22-wmt22* | 23 | 0.772 | 28 | 0.880 | 17 | 0.973 | 37 | 0.839 | 10 | 0.940 |
| *cometoid22-wmt21** | 24 | 0.768 | 30 | 0.871 | 19 | 0.973 | 38 | 0.832 | 13 | 0.929 |
| docWMT22CometDA | 25 | 0.768 | 18 | 0.904 | 2 | 0.990 | 17 | 0.922 | 17 | 0.907 |
| docWMT22CometKiwiDA* | 26 | 0.767 | 21 | 0.900 | 21 | 0.970 | 26 | 0.906 | 5 | 0.965 |
| Calibri-COMET22 | 27 | 0.767 | 15 | 0.904 | 23 | 0.963 | 13 | 0.930 | 29 | 0.863 |
| Calibri-COMET22-QE* | 28 | 0.755 | 34 | 0.863 | 11 | 0.978 | 40 | 0.778 | 12 | 0.934 |
| YiSi-1 | 29 | 0.754 | 33 | 0.871 | 31 | 0.925 | 18 | 0.917 | 32 | 0.823 |
| MS-COMET-QE-22* | 30 | 0.744 | 32 | 0.871 | 24 | 0.959 | 43 | 0.721 | 20 | 0.901 |
| prismRef | 31 | 0.744 | 37 | 0.851 | 33 | 0.920 | 3 | 0.956 | 40 | 0.762 |
| mre-score-labse-regular | 32 | 0.743 | 26 | 0.888 | 29 | 0.942 | 2 | 0.958 | 18 | 0.903 |
| BERTscore | 33 | 0.742 | 31 | 0.871 | 38 | 0.891 | 30 | 0.895 | 33 | 0.810 |
| XLsim | 34 | 0.719 | 36 | 0.855 | 32 | 0.925 | 31 | 0.887 | 36 | 0.796 |
| f200spBLEU | 35 | 0.704 | 40 | 0.819 | 34 | 0.919 | 39 | 0.805 | 39 | 0.772 |
| MEE4 | 36 | 0.704 | 39 | 0.823 | 41 | 0.861 | 32 | 0.879 | 41 | 0.743 |
| tokengram_F | 37 | 0.703 | 42 | 0.815 | 43 | 0.858 | 33 | 0.878 | 37 | 0.795 |
| embed_llama | 38 | 0.701 | 38 | 0.831 | 42 | 0.861 | 36 | 0.841 | 38 | 0.785 |
| BLEU | 39 | 0.696 | 41 | 0.815 | 37 | 0.917 | 42 | 0.769 | 42 | 0.734 |
| chrF | 40 | 0.694 | 43 | 0.795 | 40 | 0.866 | 41 | 0.776 | 34 | 0.809 |
| eBLEU | 41 | 0.692 | 35 | 0.859 | 35 | 0.918 | 20 | 0.911 | 43 | 0.727 |
| Random-sysname* | 42 | 0.529 | 44 | 0.578 | 44 | 0.357 | 44 | 0.209 | 44 | 0.093 |
| prismSrc* | 43 | 0.455 | 45 | 0.386 | 45 | -0.327 | 45 | -0.017 | 45 | -0.406 |
| *HuaweiTSC_EE_Metric* | – | – | 19 | 0.900 | 39 | 0.878 | 27 | 0.903 | 22 | 0.894 |
| *slide** | – | – | 14 | 0.904 | 15 | 0.975 | 10 | 0.938 | 24 | 0.890 |

Table 16: Results for all metrics on system-level tasks for main language pairs. Rows are sorted by the overall average correlation across all 10 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

| lang:<br>corr_fcn:<br>metric | en→de<br>pearson<br>task3 | | en→de<br>acc-t<br>task4 | | he→en<br>pearson<br>task6 | | he→en<br>acc-t<br>task7 | | zh→en<br>pearson<br>task9 | | zh→en<br>acc-t<br>task10 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| XCOMET-Ensemble | 2 | 0.695 | 3 | 0.604 | **1** | **0.556** | **1** | **0.586** | 2 | 0.650 | 2 | 0.543 |
| *XCOMET-XXL* | **1** | **0.695** | 4 | 0.603 | 2 | 0.556 | 4 | 0.577 | 5 | 0.627 | 3 | 0.541 |
| *MetricX-23-QE-b** | 5 | 0.628 | **1** | **0.606** | 6 | 0.529 | 3 | 0.580 | **1** | **0.661** | 4 | 0.539 |
| *XCOMET-XL* | 3 | 0.680 | 6 | 0.601 | 5 | 0.536 | 7 | 0.568 | 7 | 0.624 | 9 | 0.531 |
| *MetricX-23-QE-c** | 11 | 0.525 | 11 | 0.581 | 7 | 0.526 | 6 | 0.576 | 9 | 0.581 | **1** | **0.545** |
| *MetricX-23-b* | 9 | 0.566 | 2 | 0.604 | 4 | 0.537 | 2 | 0.581 | 8 | 0.612 | 6 | 0.535 |
| XCOMET-QE-Ensemble* | 4 | 0.679 | 8 | 0.588 | 9 | 0.498 | 10 | 0.554 | 4 | 0.647 | 7 | 0.533 |
| MetricX-23 | 7 | 0.585 | 5 | 0.603 | 3 | 0.548 | 5 | 0.577 | 6 | 0.625 | 8 | 0.531 |
| GEMBA-MQM* | 16 | 0.502 | 17 | 0.572 | 13 | 0.401 | 9 | 0.564 | 16 | 0.449 | 14 | 0.522 |
| MetricX-23-QE* | 6 | 0.626 | 7 | 0.596 | 8 | 0.520 | 8 | 0.564 | 3 | 0.647 | 11 | 0.527 |
| *cometoid22-wmt23** | 20 | 0.448 | 9 | 0.586 | 16 | 0.397 | 15 | 0.544 | 19 | 0.439 | 15 | 0.520 |
| mbr-metricx-qe* | 8 | 0.571 | 10 | 0.584 | 12 | 0.411 | 11 | 0.553 | 13 | 0.489 | 5 | 0.537 |
| *CometKiwi-XXL** | 28 | 0.417 | 13 | 0.578 | 19 | 0.390 | 13 | 0.550 | 24 | 0.390 | 10 | 0.528 |
| *CometKiwi-XL** | 21 | 0.446 | 18 | 0.571 | 22 | 0.384 | 18 | 0.533 | 21 | 0.430 | 13 | 0.522 |
| MaTESe | 10 | 0.554 | 30 | 0.528 | 10 | 0.459 | 12 | 0.550 | 11 | 0.511 | 34 | 0.479 |
| CometKiwi* | 18 | 0.475 | 19 | 0.569 | 20 | 0.387 | 14 | 0.544 | 18 | 0.442 | 12 | 0.525 |
| COMET | 25 | 0.432 | 15 | 0.574 | 14 | 0.401 | 19 | 0.532 | 22 | 0.396 | 19 | 0.514 |
| *MetricX-23-c* | 15 | 0.508 | 27 | 0.539 | 31 | 0.313 | 20 | 0.531 | 27 | 0.371 | 21 | 0.507 |
| *instructscore* | 12 | 0.519 | 20 | 0.563 | 11 | 0.458 | 17 | 0.536 | 12 | 0.499 | 40 | 0.459 |
| BLEURT-20 | 17 | 0.484 | 16 | 0.572 | 24 | 0.382 | 24 | 0.519 | 26 | 0.378 | 16 | 0.518 |
| KG-BERTScore* | 19 | 0.451 | 23 | 0.556 | 23 | 0.382 | 16 | 0.537 | 20 | 0.430 | 17 | 0.516 |
| sescoreX | 13 | 0.519 | 21 | 0.563 | 21 | 0.385 | 33 | 0.484 | 10 | 0.536 | 24 | 0.499 |
| cometoid22-wmt22* | 23 | 0.441 | 14 | 0.578 | 26 | 0.365 | 25 | 0.515 | 14 | 0.479 | 18 | 0.515 |
| *cometoid22-wmt21** | 26 | 0.428 | 12 | 0.581 | 27 | 0.360 | 26 | 0.515 | 15 | 0.458 | 20 | 0.514 |
| docWMT22CometDA | 30 | 0.394 | 22 | 0.559 | 28 | 0.339 | 31 | 0.497 | 29 | 0.353 | 28 | 0.493 |
| docWMT22CometKiwiDA* | 22 | 0.444 | 24 | 0.547 | 33 | 0.286 | 32 | 0.489 | 25 | 0.387 | 27 | 0.493 |
| Calibri-COMET22 | 29 | 0.413 | 34 | 0.522 | 15 | 0.401 | 27 | 0.515 | 23 | 0.396 | 36 | 0.474 |
| Calibri-COMET22-QE* | 24 | 0.441 | 41 | 0.483 | 18 | 0.395 | 29 | 0.506 | 17 | 0.443 | 29 | 0.491 |
| YiSi-1 | 31 | 0.366 | 26 | 0.542 | 17 | 0.395 | 21 | 0.529 | 30 | 0.290 | 22 | 0.504 |
| MS-COMET-QE-22* | 33 | 0.310 | 25 | 0.546 | 32 | 0.295 | 30 | 0.498 | 28 | 0.367 | 26 | 0.498 |
| prismRef | 14 | 0.516 | 38 | 0.518 | 30 | 0.319 | 22 | 0.528 | 33 | 0.183 | 23 | 0.504 |
| mre-score-labse-regular | 41 | 0.111 | 28 | 0.530 | 25 | 0.378 | 23 | 0.522 | 35 | 0.145 | 32 | 0.481 |
| BERTscore | 32 | 0.325 | 31 | 0.528 | 29 | 0.335 | 28 | 0.515 | 31 | 0.236 | 25 | 0.499 |
| XLsim | 35 | 0.239 | 32 | 0.527 | 35 | 0.233 | 34 | 0.480 | 37 | 0.111 | 39 | 0.464 |
| f200spBLEU | 36 | 0.237 | 33 | 0.526 | 36 | 0.230 | 37 | 0.447 | 38 | 0.108 | 35 | 0.476 |
| MEE4 | 39 | 0.202 | 29 | 0.529 | 34 | 0.256 | 41 | 0.441 | 39 | 0.105 | 33 | 0.480 |
| tokengram_F | 38 | 0.227 | 35 | 0.520 | 37 | 0.226 | 35 | 0.461 | 41 | 0.060 | 31 | 0.485 |
| embed_llama | 34 | 0.250 | 40 | 0.483 | 40 | 0.215 | 42 | 0.430 | 34 | 0.161 | 41 | 0.447 |
| BLEU | 40 | 0.192 | 36 | 0.520 | 39 | 0.220 | 39 | 0.442 | 36 | 0.119 | 38 | 0.472 |
| chrF | 37 | 0.232 | 37 | 0.519 | 38 | 0.221 | 36 | 0.460 | 40 | 0.063 | 30 | 0.485 |
| eBLEU | 43 | -0.011 | 39 | 0.512 | 42 | 0.131 | 38 | 0.445 | 43 | -0.084 | 37 | 0.473 |
| Random-sysname* | 42 | 0.064 | 43 | 0.409 | 43 | 0.041 | 43 | 0.428 | 42 | 0.018 | 43 | 0.381 |
| prismSrc* | 27 | 0.425 | 42 | 0.426 | 41 | 0.140 | 40 | 0.441 | 32 | 0.223 | 42 | 0.421 |

Table 17: Results for all metrics on segment-level tasks for main language pairs. Rows are sorted by the overall average correlation across all 10 tasks (leftmost column in Table 16). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

| metric | | avg corr | p-values | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| XCOMET-Ensemble | 1 | 0.825 | . | 01 | 01 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| XCOMET-QE-Ensemble* | 2 | 0.808 | . | . | 46 | 20 | 26 | 00 | 00 | 00 | 01 | 01 | 01 | 03 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| MetricX-23 | 2 | 0.808 | . | . | . | 24 | 25 | 03 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| GEMBA-MQM* | 2 | 0.802 | . | . | . | . | 43 | 03 | 00 | 00 | 00 | 00 | 00 | 02 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| MetricX-23-QE* | 2 | 0.800 | . | . | . | . | . | 13 | 07 | 05 | 03 | 00 | 06 | 02 | 00 | 00 | 00 | 00 | 01 | 01 | 00 | 00 | 02 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| mbr-metricx-qe* | 3 | 0.788 | . | . | . | . | . | . | 31 | 24 | 17 | 10 | 16 | 09 | 02 | 02 | 00 | 00 | 00 | 03 | 00 | 02 | 01 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| MaTESe | 3 | 0.782 | . | . | . | . | . | . | . | 48 | 38 | 26 | 19 | 24 | 12 | 09 | 04 | 06 | 03 | 03 | 00 | 01 | 03 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| CometKiwi* | 3 | 0.782 | . | . | . | . | . | . | . | . | 39 | 25 | 26 | 23 | 04 | 07 | 01 | 02 | 02 | 02 | 00 | 00 | 01 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| COMET | 3 | 0.779 | . | . | . | . | . | . | . | . | . | 22 | 34 | 25 | 23 | 01 | 19 | 11 | 11 | 01 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| BLEURT-20 | 3 | 0.776 | . | . | . | . | . | . | . | . | . | . | 46 | 34 | 35 | 10 | 20 | 16 | 13 | 04 | 02 | 00 | 00 | 01 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| KG-BERTScore* | 3 | 0.774 | . | . | . | . | . | . | . | . | . | . | . | 43 | 49 | 24 | 29 | 32 | 16 | 08 | 00 | 04 | 07 | 02 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| sescoreX | 3 | 0.772 | . | . | . | . | . | . | . | . | . | . | . | . | 49 | 30 | 37 | 31 | 18 | 06 | 08 | 03 | 04 | 04 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| cometoid22-wmt22* | 4 | 0.772 | . | . | . | . | . | . | . | . | . | . | . | . | . | 34 | 22 | 22 | 07 | 14 | 03 | 04 | 07 | 01 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| docWMT22CometDA | 4 | 0.768 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 51 | 44 | 24 | 10 | 03 | 03 | 03 | 04 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| docWMT22CometKiwiDA* | 4 | 0.767 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 48 | 14 | 20 | 07 | 09 | 12 | 03 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| Calibri-COMET22 | 4 | 0.767 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 17 | 23 | 10 | 16 | 11 | 01 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| Calibri-COMET22-QE* | 4 | 0.755 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 45 | 30 | 36 | 30 | 18 | 07 | 01 | 04 | 02 | 01 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| YiSi-1 | 4 | 0.754 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 30 | 13 | 22 | 31 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| MS-COMET-QE-22* | 5 | 0.744 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 52 | 49 | 43 | 12 | 01 | 02 | 02 | 02 | 01 | 00 | 00 | 00 | 00 | 00 | 00 |
| prismRef | 5 | 0.744 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 44 | 44 | 00 | 00 | 01 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| mre-score-labse-regular | 5 | 0.743 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 49 | 06 | 01 | 04 | 01 | 00 | 00 | 00 | 00 | 00 | 00 | 00 | 00 |
| BERTscore | 5 | 0.742 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 18 | 03 | 07 | 05 | 01 | 02 | 02 | 00 | 00 | 00 | 00 | 00 |
| XLsim | 6 | 0.719 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 04 | 10 | 01 | 06 | 01 | 01 | 00 | 00 | 00 | 00 | 00 |
| f200spBLEU | 7 | 0.704 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 51 | 48 | 39 | 06 | 13 | 12 | 00 | 00 | 00 | 00 |
| MEE4 | 7 | 0.704 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 46 | 46 | 33 | 23 | 16 | 00 | 00 | 00 | 00 |
| tokengram_F | 7 | 0.703 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 45 | 22 | 15 | 11 | 00 | 00 | 00 | 00 |
| embed_llama | 7 | 0.701 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 34 | 30 | 29 | 00 | 00 | 00 | 00 |
| BLEU | 7 | 0.696 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 42 | 35 | 00 | 00 | 00 | 00 |
| chrF | 7 | 0.694 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 43 | 00 | 00 | 00 | 00 |
| eBLEU | 7 | 0.692 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 00 | 00 | 00 | 00 |
| Random-sysname* | 8 | 0.529 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 04 | 00 | 00 |
| prismSrc* | 9 | 0.455 | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . | 00 | 00 |

Table 18: Results of pairwise metric significance tests for primary submissions using permutation resampling. Each value gives the 100 $\times$ estimated probability of the null hypothesis that the average correlation of the metric in the current row is $\leq$ the average correlation of the metric in the current column. Starred metrics are reference-free, and underlined metrics are baselines.

## B Significance comparisons for main results

Table 18 contains the results of pairwise comparisons for the results in Table 1.

## C Correlations with WMT DA-SQM for all metrics

Tables 19 to 27 give correlations with WMT direct assessment (DA-SQM) scores on all 8 translation directions for which those scores are available. In all cases, reference A was used, and no additional metrics were available to be scored by the metrics. We evaluate metrics on a task setup similar to that of Table 7: one system-level pairwise accuracy task involving all languages (with a weight of 8), and system-level Pearson, segment-level Pearson, and segment-level $\text{acc}^*_{eq}$ tasks for each translation direction (24 tasks in total, each with a weight of 1). Each table shows overall average correlation, along with the results for the tasks for one translation direction. Metrics that did not participate in all tasks do not have an average correlation, and are displayed at the end of each table.

We wish to emphasize that the DA+SQM is considerably noisier than MQM. This increased variability may influence the outcomes observed in the following spotlight evaluation. Consequently, readers should exercise considerable caution when drawing conclusions from these results.

| lang:<br>corr_fcn:<br>metric | avg-corr | | cs→uk,de→en,en→cs,en→de,en→ja,en→zh,ja→en,zh→en<br>accuracy<br>task1 | |
|---|---|---|---|---|
| *CometKiwi-XXL** | **1** | **0.798** | **1** | **0.912** |
| *CometKiwi-XL** | 2 | 0.795 | 2 | 0.905 |
| COMET | 3 | 0.787 | 8 | 0.890 |
| CometKiwi* | 4 | 0.787 | 10 | 0.887 |
| *cometoid22-wmt23** | 5 | 0.786 | 6 | 0.897 |
| KG-BERTScore* | 6 | 0.784 | 12 | 0.884 |
| *MetricX-23-QE-c** | 7 | 0.780 | 9 | 0.887 |
| BLEURT-20 | 8 | 0.778 | 15 | 0.880 |
| *MetricX-23-QE-b** | 9 | 0.777 | 14 | 0.880 |
| cometoid22-wmt22* | 10 | 0.776 | 13 | 0.884 |
| *MetricX-23-c* | 11 | 0.775 | 5 | 0.898 |
| *cometoid22-wmt21** | 12 | 0.774 | 11 | 0.885 |
| XCOMET-Ensemble | 13 | 0.774 | 20 | 0.870 |
| *MetricX-23-b* | 14 | 0.768 | 17 | 0.873 |
| MetricX-23-QE* | 15 | 0.768 | 19 | 0.870 |
| MS-COMET-QE-22* | 16 | 0.767 | 16 | 0.879 |
| XCOMET-QE-Ensemble* | 17 | 0.766 | 18 | 0.871 |
| MetricX-23 | 18 | 0.762 | 22 | 0.863 |
| YiSi-1 | 19 | 0.749 | 25 | 0.832 |
| *XCOMET-XL* | 20 | 0.748 | 24 | 0.860 |
| XLsim | 21 | 0.745 | 26 | 0.831 |
| *XCOMET-XXL* | 22 | 0.743 | 21 | 0.866 |
| GEMBA-MQM* | 23 | 0.739 | 4 | 0.899 |
| prismRef | 24 | 0.736 | 27 | 0.808 |
| mre-score-labse-regular | 25 | 0.734 | 23 | 0.861 |
| BERTscore | 26 | 0.732 | 28 | 0.799 |
| tokengram_F | 27 | 0.714 | 30 | 0.786 |
| chrF | 28 | 0.712 | 33 | 0.777 |
| f200spBLEU | 29 | 0.708 | 29 | 0.786 |
| embed_llama | 30 | 0.701 | 32 | 0.778 |
| eBLEU | 31 | 0.694 | 31 | 0.781 |
| BLEU | 32 | 0.660 | 34 | 0.770 |
| Random-sysname* | 33 | 0.537 | 35 | 0.580 |
| prismSrc* | 34 | 0.514 | 36 | 0.412 |
| *HuaweiTSC_EE_Metric* | – | – | 7 | 0.892 |
| *slide** | – | – | 3 | 0.902 |

Table 19: Correlations with WMT DA-SQM scores for all metrics on all-pairs data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

| lang:<br>corr_fcn:<br>metric | avg-corr | | cs→uk<br>pearson<br>task1 | | cs→uk<br>pearson<br>task2 | | cs→uk<br>acc-t<br>task3 | |
|---|---|---|---|---|---|---|---|---|
| *CometKiwi-XXL** | **1** | **0.798** | 12 | 0.889 | 4 | 0.462 | 6 | 0.555 |
| *CometKiwi-XL** | 2 | 0.795 | 18 | 0.866 | 15 | 0.412 | 10 | 0.548 |
| COMET | 3 | 0.787 | 5 | 0.899 | 6 | 0.454 | 8 | 0.553 |
| CometKiwi* | 4 | 0.787 | 23 | 0.788 | 8 | 0.429 | 13 | 0.536 |
| *cometoid22-wmt23** | 5 | 0.786 | 7 | 0.898 | 11 | 0.420 | 15 | 0.534 |
| KG-BERTScore* | 6 | 0.784 | 24 | 0.788 | 9 | 0.429 | 16 | 0.530 |
| *MetricX-23-QE-c** | 7 | 0.780 | 3 | 0.920 | 2 | 0.502 | 9 | 0.553 |
| BLEURT-20 | 8 | 0.778 | 2 | 0.926 | 7 | 0.443 | 12 | 0.538 |
| *MetricX-23-QE-b** | 9 | 0.777 | 6 | 0.898 | 16 | 0.410 | 4 | 0.559 |
| cometoid22-wmt22* | 10 | 0.776 | 19 | 0.851 | 19 | 0.403 | 19 | 0.528 |
| *MetricX-23-c* | 11 | 0.775 | **1** | **0.932** | **1** | **0.523** | 5 | 0.558 |
| *cometoid22-wmt21** | 12 | 0.774 | 21 | 0.822 | 14 | 0.414 | 23 | 0.521 |
| XCOMET-Ensemble | 13 | 0.774 | 8 | 0.897 | 3 | 0.482 | 3 | 0.560 |
| *MetricX-23-b* | 14 | 0.768 | 13 | 0.888 | 17 | 0.410 | **1** | **0.568** |
| MetricX-23-QE* | 15 | 0.768 | 11 | 0.889 | 21 | 0.382 | 7 | 0.555 |
| MS-COMET-QE-22* | 16 | 0.767 | 20 | 0.851 | 23 | 0.322 | 24 | 0.519 |
| XCOMET-QE-Ensemble* | 17 | 0.766 | 17 | 0.873 | 5 | 0.462 | 11 | 0.540 |
| MetricX-23 | 18 | 0.762 | 15 | 0.879 | 20 | 0.395 | 2 | 0.567 |
| YiSi-1 | 19 | 0.749 | 26 | 0.753 | 25 | 0.315 | 20 | 0.526 |
| *XCOMET-XL* | 20 | 0.748 | 14 | 0.882 | 10 | 0.423 | 18 | 0.529 |
| XLsim | 21 | 0.745 | 22 | 0.792 | 24 | 0.318 | 21 | 0.526 |
| *XCOMET-XXL* | 22 | 0.743 | 9 | 0.897 | 18 | 0.407 | 33 | 0.436 |
| GEMBA-MQM* | 23 | 0.739 | 4 | 0.913 | 12 | 0.419 | 34 | 0.323 |
| prismRef | 24 | 0.736 | 27 | 0.694 | 22 | 0.372 | 17 | 0.530 |
| mre-score-labse-regular | 25 | 0.734 | 25 | 0.772 | 13 | 0.417 | 14 | 0.534 |
| BERTscore | 26 | 0.732 | 32 | 0.544 | 26 | 0.292 | 22 | 0.524 |
| tokengram_F | 27 | 0.714 | 30 | 0.626 | 28 | 0.268 | 25 | 0.518 |
| chrF | 28 | 0.712 | 29 | 0.637 | 27 | 0.273 | 26 | 0.517 |
| f200spBLEU | 29 | 0.708 | 28 | 0.676 | 30 | 0.221 | 28 | 0.504 |
| embed_llama | 30 | 0.701 | 34 | 0.511 | 33 | 0.157 | 30 | 0.492 |
| eBLEU | 31 | 0.694 | 33 | 0.512 | 31 | 0.188 | 27 | 0.511 |
| BLEU | 32 | 0.660 | 31 | 0.548 | 32 | 0.184 | 31 | 0.480 |
| Random-sysname* | 33 | 0.537 | 35 | 0.343 | 34 | 0.047 | 32 | 0.469 |
| prismSrc* | 34 | 0.514 | 36 | -0.236 | 29 | 0.261 | 29 | 0.495 |
| *HuaweiTSC_EE_Metric* | – | – | 10 | 0.893 | – | – | – | – |
| *slide** | – | – | 16 | 0.877 | – | – | – | – |

Table 20: Correlations with WMT DA-SQM scores for all metrics on cs→uk data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

| lang:<br>corr_fcn:<br>metric | avg-corr | | de→en<br>pearson<br>task1 | | de→en<br>pearson<br>task2 | | de→en<br>acc-t<br>task3 | |
|---|---|---|---|---|---|---|---|---|
| *CometKiwi-XXL** | **1** | **0.798** | 15 | 0.931 | 14 | 0.411 | 11 | 0.571 |
| *CometKiwi-XL** | 2 | 0.795 | 12 | 0.934 | 17 | 0.402 | 13 | 0.569 |
| COMET | 3 | 0.787 | 6 | 0.953 | 2 | 0.480 | 2 | 0.584 |
| CometKiwi* | 4 | 0.787 | 14 | 0.933 | 9 | 0.447 | 16 | 0.559 |
| *cometoid22-wmt23** | 5 | 0.786 | 17 | 0.913 | 3 | 0.471 | 12 | 0.571 |
| KG-BERTScore* | 6 | 0.784 | 13 | 0.933 | 7 | 0.447 | 20 | 0.553 |
| *MetricX-23-QE-c** | 7 | 0.780 | 30 | 0.835 | 8 | 0.447 | 7 | 0.574 |
| BLEURT-20 | 8 | 0.778 | 3 | 0.965 | **1** | **0.486** | 5 | 0.578 |
| *MetricX-23-QE-b** | 9 | 0.777 | 21 | 0.893 | 12 | 0.425 | 4 | 0.579 |
| cometoid22-wmt22* | 10 | 0.776 | 23 | 0.881 | 6 | 0.449 | 17 | 0.558 |
| *MetricX-23-c* | 11 | 0.775 | 9 | 0.944 | 27 | 0.298 | 24 | 0.544 |
| *cometoid22-wmt21** | 12 | 0.774 | 26 | 0.856 | 10 | 0.437 | 19 | 0.556 |
| XCOMET-Ensemble | 13 | 0.774 | 28 | 0.842 | 15 | 0.408 | 9 | 0.573 |
| *MetricX-23-b* | 14 | 0.768 | 27 | 0.850 | 18 | 0.389 | **1** | **0.590** |
| MetricX-23-QE* | 15 | 0.768 | 24 | 0.876 | 13 | 0.418 | 8 | 0.574 |
| MS-COMET-QE-22* | 16 | 0.767 | 29 | 0.841 | 32 | 0.256 | 23 | 0.545 |
| XCOMET-QE-Ensemble* | 17 | 0.766 | 34 | 0.813 | 19 | 0.385 | 18 | 0.556 |
| MetricX-23 | 18 | 0.762 | 31 | 0.831 | 20 | 0.382 | 3 | 0.584 |
| YiSi-1 | 19 | 0.749 | **1** | **0.970** | 5 | 0.451 | 10 | 0.572 |
| *XCOMET-XL* | 20 | 0.748 | 35 | 0.780 | 22 | 0.341 | 25 | 0.544 |
| XLsim | 21 | 0.745 | 8 | 0.947 | 23 | 0.340 | 15 | 0.560 |
| *XCOMET-XXL* | 22 | 0.743 | 32 | 0.828 | 21 | 0.375 | 31 | 0.517 |
| GEMBA-MQM* | 23 | 0.739 | 10 | 0.938 | 4 | 0.463 | 34 | 0.426 |
| prismRef | 24 | 0.736 | 4 | 0.963 | 16 | 0.403 | 14 | 0.565 |
| mre-score-labse-regular | 25 | 0.734 | 16 | 0.916 | 34 | 0.121 | 26 | 0.540 |
| BERTscore | 26 | 0.732 | 2 | 0.969 | 11 | 0.434 | 6 | 0.576 |
| tokengram_F | 27 | 0.714 | 22 | 0.891 | 25 | 0.319 | 21 | 0.551 |
| chrF | 28 | 0.712 | 25 | 0.860 | 24 | 0.328 | 22 | 0.550 |
| f200spBLEU | 29 | 0.708 | 19 | 0.904 | 28 | 0.291 | 27 | 0.539 |
| embed_llama | 30 | 0.701 | 18 | 0.913 | 29 | 0.275 | 30 | 0.525 |
| eBLEU | 31 | 0.694 | 5 | 0.954 | 33 | 0.207 | 28 | 0.538 |
| BLEU | 32 | 0.660 | 20 | 0.897 | 31 | 0.270 | 29 | 0.534 |
| Random-sysname* | 33 | 0.537 | 37 | 0.185 | 35 | 0.044 | 33 | 0.472 |
| prismSrc* | 34 | 0.514 | 36 | 0.449 | 30 | 0.273 | 32 | 0.502 |
| *HuaweiTSC_EE_Metric* | – | – | 7 | 0.950 | – | – | – | – |
| *slide** | – | – | 11 | 0.934 | – | – | – | – |
| MaTESe | – | – | 33 | 0.816 | 26 | 0.308 | 35 | 0.373 |

Table 21: Correlations with WMT DA-SQM scores for all metrics on de→en data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

| lang:<br>corr_fcn:<br>metric | avg-corr | | en→cs<br>pearson<br>task1 | | en→cs<br>pearson<br>task2 | | en→cs<br>acc-t<br>task3 | |
|---|---|---|---|---|---|---|---|---|
| *CometKiwi-XXL** | **1** | **0.798** | **1** | **0.922** | 7 | 0.367 | 5 | 0.548 |
| *CometKiwi-XL** | 2 | 0.795 | 5 | 0.897 | 6 | 0.369 | 7 | 0.541 |
| COMET | 3 | 0.787 | 14 | 0.865 | 4 | 0.377 | 10 | 0.524 |
| CometKiwi* | 4 | 0.787 | 22 | 0.790 | 13 | 0.350 | 11 | 0.518 |
| *cometoid22-wmt23** | 5 | 0.786 | 13 | 0.865 | 11 | 0.352 | 16 | 0.507 |
| KG-BERTScore* | 6 | 0.784 | 21 | 0.790 | 12 | 0.350 | 17 | 0.507 |
| *MetricX-23-QE-c** | 7 | 0.780 | 6 | 0.893 | 3 | 0.391 | 8 | 0.540 |
| BLEURT-20 | 8 | 0.778 | 20 | 0.793 | 5 | 0.373 | 12 | 0.510 |
| *MetricX-23-QE-b** | 9 | 0.777 | 10 | 0.881 | 18 | 0.338 | 2 | 0.551 |
| cometoid22-wmt22* | 10 | 0.776 | 17 | 0.825 | 16 | 0.341 | 18 | 0.506 |
| *MetricX-23-c* | 11 | 0.775 | 23 | 0.750 | 19 | 0.316 | 13 | 0.510 |
| *cometoid22-wmt21** | 12 | 0.774 | 18 | 0.824 | 17 | 0.340 | 14 | 0.508 |
| XCOMET-Ensemble | 13 | 0.774 | 3 | 0.903 | **1** | **0.402** | 6 | 0.543 |
| *MetricX-23-b* | 14 | 0.768 | 11 | 0.880 | 15 | 0.344 | **1** | **0.552** |
| MetricX-23-QE* | 15 | 0.768 | 12 | 0.878 | 14 | 0.348 | 4 | 0.549 |
| MS-COMET-QE-22* | 16 | 0.767 | 19 | 0.797 | 21 | 0.286 | 21 | 0.497 |
| XCOMET-QE-Ensemble* | 17 | 0.766 | 2 | 0.908 | 2 | 0.395 | 9 | 0.528 |
| MetricX-23 | 18 | 0.762 | 7 | 0.891 | 9 | 0.361 | 3 | 0.550 |
| YiSi-1 | 19 | 0.749 | 26 | 0.568 | 24 | 0.245 | 24 | 0.492 |
| *XCOMET-XL* | 20 | 0.748 | 4 | 0.898 | 8 | 0.362 | 15 | 0.507 |
| XLsim | 21 | 0.745 | 25 | 0.627 | 23 | 0.259 | 20 | 0.503 |
| *XCOMET-XXL* | 22 | 0.743 | 8 | 0.890 | 10 | 0.353 | 33 | 0.439 |
| GEMBA-MQM* | 23 | 0.739 | 16 | 0.852 | 20 | 0.309 | 34 | 0.327 |
| prismRef | 24 | 0.736 | 27 | 0.557 | 22 | 0.265 | 22 | 0.495 |
| mre-score-labse-regular | 25 | 0.734 | 24 | 0.718 | 33 | 0.130 | 19 | 0.504 |
| BERTscore | 26 | 0.732 | 30 | 0.480 | 25 | 0.228 | 23 | 0.493 |
| tokengram_F | 27 | 0.714 | 34 | 0.409 | 26 | 0.203 | 26 | 0.481 |
| chrF | 28 | 0.712 | 33 | 0.450 | 27 | 0.201 | 27 | 0.480 |
| f200spBLEU | 29 | 0.708 | 29 | 0.496 | 28 | 0.199 | 29 | 0.475 |
| embed_llama | 30 | 0.701 | 32 | 0.466 | 30 | 0.172 | 28 | 0.476 |
| eBLEU | 31 | 0.694 | 31 | 0.467 | 32 | 0.169 | 25 | 0.483 |
| BLEU | 32 | 0.660 | 28 | 0.519 | 29 | 0.186 | 30 | 0.460 |
| Random-sysname* | 33 | 0.537 | 35 | 0.015 | 34 | 0.002 | 32 | 0.452 |
| prismSrc* | 34 | 0.514 | 36 | -0.042 | 31 | 0.171 | 31 | 0.456 |
| *HuaweiTSC_EE_Metric* | – | – | 15 | 0.862 | – | – | – | – |
| *slide** | – | – | 9 | 0.885 | – | – | – | – |

Table 22: Correlations with WMT DA-SQM scores for all metrics on en→cs data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

| lang:<br>corr_fcn:<br>metric | avg-corr | | en→de<br>pearson<br>task1 | | en→de<br>pearson<br>task2 | | en→de<br>acc-t<br>task3 | |
|---|---|---|---|---|---|---|---|---|
| *CometKiwi-XXL** | **1** | **0.798** | 13 | 0.972 | 4 | 0.506 | **1** | **0.595** |
| *CometKiwi-XL** | 2 | 0.795 | 5 | 0.984 | 3 | 0.512 | 3 | 0.589 |
| COMET | 3 | 0.787 | 21 | 0.953 | 5 | 0.496 | 5 | 0.588 |
| CometKiwi* | 4 | 0.787 | **1** | **0.990** | **1** | **0.537** | 6 | 0.586 |
| *cometoid22-wmt23** | 5 | 0.786 | 26 | 0.944 | 6 | 0.491 | 12 | 0.580 |
| KG-BERTScore* | 6 | 0.784 | 3 | 0.990 | 2 | 0.523 | 17 | 0.578 |
| *MetricX-23-QE-c** | 7 | 0.780 | 39 | 0.859 | 12 | 0.465 | 19 | 0.576 |
| BLEURT-20 | 8 | 0.778 | 25 | 0.945 | 15 | 0.452 | 18 | 0.577 |
| *MetricX-23-QE-b** | 9 | 0.777 | 33 | 0.910 | 19 | 0.437 | 4 | 0.588 |
| cometoid22-wmt22* | 10 | 0.776 | 32 | 0.911 | 16 | 0.447 | 21 | 0.575 |
| *MetricX-23-c* | 11 | 0.775 | 2 | 0.990 | 8 | 0.482 | 13 | 0.580 |
| *cometoid22-wmt21** | 12 | 0.774 | 34 | 0.905 | 20 | 0.433 | 22 | 0.574 |
| XCOMET-Ensemble | 13 | 0.774 | 38 | 0.861 | 25 | 0.399 | 20 | 0.576 |
| *MetricX-23-b* | 14 | 0.768 | 35 | 0.896 | 30 | 0.377 | 8 | 0.583 |
| MetricX-23-QE* | 15 | 0.768 | 37 | 0.867 | 18 | 0.443 | 11 | 0.582 |
| MS-COMET-QE-22* | 16 | 0.767 | 28 | 0.942 | 33 | 0.371 | 28 | 0.558 |
| XCOMET-QE-Ensemble* | 17 | 0.766 | 41 | 0.849 | 29 | 0.382 | 27 | 0.564 |
| MetricX-23 | 18 | 0.762 | 40 | 0.855 | 28 | 0.389 | 9 | 0.582 |
| YiSi-1 | 19 | 0.749 | 6 | 0.980 | 13 | 0.456 | 23 | 0.571 |
| *XCOMET-XL* | 20 | 0.748 | 42 | 0.845 | 34 | 0.365 | 32 | 0.552 |
| XLsim | 21 | 0.745 | 7 | 0.979 | 27 | 0.391 | 25 | 0.566 |
| *XCOMET-XXL* | 22 | 0.743 | 36 | 0.868 | 24 | 0.399 | 39 | 0.525 |
| GEMBA-MQM* | 23 | 0.739 | 17 | 0.961 | 7 | 0.488 | 42 | 0.434 |
| prismRef | 24 | 0.736 | 16 | 0.963 | 37 | 0.321 | 36 | 0.544 |
| mre-score-labse-regular | 25 | 0.734 | 30 | 0.927 | 42 | 0.144 | 35 | 0.548 |
| BERTscore | 26 | 0.732 | 12 | 0.973 | 23 | 0.417 | 24 | 0.567 |
| tokengram_F | 27 | 0.714 | 27 | 0.943 | 32 | 0.371 | 30 | 0.556 |
| chrF | 28 | 0.712 | 24 | 0.945 | 31 | 0.374 | 31 | 0.553 |
| f200spBLEU | 29 | 0.708 | 14 | 0.970 | 36 | 0.324 | 29 | 0.557 |
| embed_llama | 30 | 0.701 | 23 | 0.951 | 35 | 0.348 | 34 | 0.550 |
| eBLEU | 31 | 0.694 | 31 | 0.920 | 41 | 0.159 | 37 | 0.542 |
| BLEU | 32 | 0.660 | 18 | 0.958 | 38 | 0.275 | 38 | 0.541 |
| Random-sysname* | 33 | 0.537 | 44 | 0.278 | 43 | 0.075 | 41 | 0.482 |
| prismSrc* | 34 | 0.514 | 45 | -0.364 | 40 | 0.190 | 40 | 0.485 |
| *HuaweiTSC_EE_Metric* | – | – | 10 | 0.975 | – | – | – | – |
| *instructscore* | – | – | 8 | 0.977 | 10 | 0.473 | 15 | 0.578 |
| *slide** | – | – | 4 | 0.984 | – | – | – | – |
| Calibri-COMET22 | – | – | 22 | 0.953 | 21 | 0.425 | 7 | 0.584 |
| Calibri-COMET22-QE* | – | – | 19 | 0.957 | 17 | 0.445 | 33 | 0.551 |
| MEE4 | – | – | 15 | 0.968 | 22 | 0.421 | 26 | 0.565 |
| MaTESe | – | – | 43 | 0.791 | 39 | 0.272 | 43 | 0.375 |
| docWMT22CometDA | – | – | 29 | 0.941 | 14 | 0.454 | 2 | 0.593 |
| docWMT22CometKiwiDA* | – | – | 11 | 0.973 | 26 | 0.392 | 14 | 0.579 |
| mbr-metricx-qe* | – | – | 20 | 0.954 | 9 | 0.477 | 10 | 0.582 |
| sescoreX | – | – | 9 | 0.977 | 11 | 0.473 | 16 | 0.578 |

Table 23: Correlations with WMT DA-SQM scores for all metrics on en→de data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

| lang:<br>corr_fcn:<br>metric | avg-corr | | en→ja<br>pearson<br>task1 | | en→ja<br>pearson<br>task2 | | en→ja<br>acc-t<br>task3 | |
|---|---|---|---|---|---|---|---|---|
| *CometKiwi-XXL** | **1** | **0.798** | 3 | 0.993 | 2 | 0.527 | 2 | 0.592 |
| *CometKiwi-XL** | 2 | 0.795 | 2 | 0.993 | **1** | **0.528** | **1** | **0.593** |
| COMET | 3 | 0.787 | 12 | 0.969 | 6 | 0.462 | 11 | 0.580 |
| CometKiwi* | 4 | 0.787 | 6 | 0.984 | 4 | 0.516 | 4 | 0.588 |
| *cometoid22-wmt23** | 5 | 0.786 | 11 | 0.979 | 10 | 0.449 | 13 | 0.574 |
| KG-BERTScore* | 6 | 0.784 | 5 | 0.984 | 3 | 0.516 | 7 | 0.583 |
| *MetricX-23-QE-c** | 7 | 0.780 | 22 | 0.955 | 8 | 0.456 | 9 | 0.580 |
| BLEURT-20 | 8 | 0.778 | 4 | 0.990 | 15 | 0.417 | 15 | 0.569 |
| *MetricX-23-QE-b** | 9 | 0.777 | 21 | 0.956 | 14 | 0.428 | 3 | 0.590 |
| cometoid22-wmt22* | 10 | 0.776 | 20 | 0.960 | 11 | 0.449 | 14 | 0.569 |
| *MetricX-23-c* | 11 | 0.775 | 28 | 0.918 | 23 | 0.371 | 25 | 0.545 |
| *cometoid22-wmt21** | 12 | 0.774 | 16 | 0.964 | 12 | 0.442 | 16 | 0.568 |
| XCOMET-Ensemble | 13 | 0.774 | 26 | 0.920 | 5 | 0.470 | 6 | 0.586 |
| *MetricX-23-b* | 14 | 0.768 | 23 | 0.941 | 16 | 0.413 | 5 | 0.587 |
| MetricX-23-QE* | 15 | 0.768 | 30 | 0.898 | 17 | 0.411 | 10 | 0.580 |
| MS-COMET-QE-22* | 16 | 0.767 | 9 | 0.983 | 7 | 0.458 | 18 | 0.565 |
| XCOMET-QE-Ensemble* | 17 | 0.766 | 31 | 0.895 | 9 | 0.455 | 12 | 0.574 |
| MetricX-23 | 18 | 0.762 | 29 | 0.916 | 18 | 0.401 | 8 | 0.580 |
| YiSi-1 | 19 | 0.749 | 7 | 0.984 | 21 | 0.382 | 20 | 0.561 |
| *XCOMET-XL* | 20 | 0.748 | 34 | 0.821 | 19 | 0.397 | 21 | 0.558 |
| XLsim | 21 | 0.745 | 27 | 0.918 | 24 | 0.354 | 22 | 0.557 |
| *XCOMET-XXL* | 22 | 0.743 | 32 | 0.871 | 20 | 0.394 | 31 | 0.485 |
| GEMBA-MQM* | 23 | 0.739 | 8 | 0.983 | 13 | 0.429 | 33 | 0.389 |
| prismRef | 24 | 0.736 | 25 | 0.922 | 22 | 0.371 | 19 | 0.561 |
| mre-score-labse-regular | 25 | 0.734 | 10 | 0.979 | 31 | 0.120 | 17 | 0.566 |
| BERTscore | 26 | 0.732 | 18 | 0.962 | 26 | 0.317 | 23 | 0.550 |
| tokengram_F | 27 | 0.714 | 13 | 0.969 | 27 | 0.227 | 24 | 0.548 |
| chrF | 28 | 0.712 | 14 | 0.966 | 28 | 0.220 | 26 | 0.543 |
| f200spBLEU | 29 | 0.708 | 19 | 0.961 | 30 | 0.190 | 29 | 0.523 |
| embed_llama | 30 | 0.701 | 15 | 0.964 | 29 | 0.212 | 28 | 0.524 |
| eBLEU | 31 | 0.694 | 24 | 0.926 | 32 | 0.073 | 30 | 0.522 |
| BLEU | 32 | 0.660 | 33 | 0.833 | 34 | 0.001 | 34 | 0.070 |
| Random-sysname* | 33 | 0.537 | 36 | 0.307 | 33 | 0.064 | 32 | 0.484 |
| prismSrc* | 34 | 0.514 | 35 | 0.764 | 25 | 0.322 | 27 | 0.530 |
| *HuaweiTSC_EE_Metric* | – | – | 17 | 0.963 | – | – | – | – |
| *slide** | – | – | **1** | **0.995** | – | – | – | – |

Table 24: Correlations with WMT DA-SQM scores for all metrics on en→ja data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

| lang:<br>corr_fcn:<br>metric | avg-corr | | en→zh<br>pearson<br>task1 | | en→zh<br>pearson<br>task2 | | en→zh<br>acc-t<br>task3 | |
|---|---|---|---|---|---|---|---|---|
| *CometKiwi-XXL** | **1** | **0.798** | 14 | 0.982 | 7 | 0.559 | 3 | 0.601 |
| *CometKiwi-XL** | 2 | 0.795 | 8 | 0.988 | 4 | 0.588 | **1** | **0.601** |
| COMET | 3 | 0.787 | 3 | 0.995 | 6 | 0.575 | 8 | 0.589 |
| CometKiwi* | 4 | 0.787 | 4 | 0.994 | **1** | **0.635** | 7 | 0.590 |
| *cometoid22-wmt23** | 5 | 0.786 | **1** | **0.997** | 5 | 0.588 | 11 | 0.584 |
| KG-BERTScore* | 6 | 0.784 | 5 | 0.994 | 2 | 0.635 | 10 | 0.584 |
| *MetricX-23-QE-c** | 7 | 0.780 | 28 | 0.913 | 18 | 0.468 | 12 | 0.582 |
| BLEURT-20 | 8 | 0.778 | 9 | 0.988 | 8 | 0.550 | 18 | 0.571 |
| *MetricX-23-QE-b** | 9 | 0.777 | 19 | 0.963 | 19 | 0.456 | 2 | 0.601 |
| cometoid22-wmt22* | 10 | 0.776 | 7 | 0.989 | 9 | 0.537 | 15 | 0.574 |
| *MetricX-23-c* | 11 | 0.775 | 24 | 0.937 | 12 | 0.507 | 23 | 0.563 |
| *cometoid22-wmt21** | 12 | 0.774 | 10 | 0.988 | 10 | 0.527 | 16 | 0.573 |
| XCOMET-Ensemble | 13 | 0.774 | 21 | 0.944 | 14 | 0.493 | 4 | 0.596 |
| *MetricX-23-b* | 14 | 0.768 | 27 | 0.926 | 23 | 0.420 | 5 | 0.595 |
| MetricX-23-QE* | 15 | 0.768 | 22 | 0.943 | 22 | 0.439 | 6 | 0.594 |
| MS-COMET-QE-22* | 16 | 0.767 | 2 | 0.996 | 3 | 0.610 | 19 | 0.570 |
| XCOMET-QE-Ensemble* | 17 | 0.766 | 30 | 0.908 | 21 | 0.450 | 14 | 0.577 |
| MetricX-23 | 18 | 0.762 | 33 | 0.885 | 24 | 0.411 | 9 | 0.588 |
| YiSi-1 | 19 | 0.749 | 15 | 0.977 | 15 | 0.493 | 20 | 0.566 |
| *XCOMET-XL* | 20 | 0.748 | 35 | 0.790 | 26 | 0.366 | 28 | 0.542 |
| XLsim | 21 | 0.745 | 12 | 0.985 | 11 | 0.524 | 17 | 0.572 |
| *XCOMET-XXL* | 22 | 0.743 | 32 | 0.885 | 25 | 0.391 | 31 | 0.517 |
| GEMBA-MQM* | 23 | 0.739 | 18 | 0.973 | 16 | 0.489 | 33 | 0.385 |
| prismRef | 24 | 0.736 | 17 | 0.975 | 13 | 0.496 | 21 | 0.564 |
| mre-score-labse-regular | 25 | 0.734 | 11 | 0.986 | 32 | 0.177 | 13 | 0.577 |
| BERTscore | 26 | 0.732 | 16 | 0.975 | 17 | 0.474 | 22 | 0.563 |
| tokengram_F | 27 | 0.714 | 20 | 0.945 | 27 | 0.343 | 24 | 0.558 |
| chrF | 28 | 0.712 | 25 | 0.934 | 29 | 0.326 | 25 | 0.550 |
| f200spBLEU | 29 | 0.708 | 31 | 0.905 | 28 | 0.327 | 26 | 0.547 |
| embed_llama | 30 | 0.701 | 26 | 0.927 | 30 | 0.297 | 27 | 0.542 |
| eBLEU | 31 | 0.694 | 29 | 0.912 | 31 | 0.210 | 29 | 0.535 |
| BLEU | 32 | 0.660 | 34 | 0.804 | 33 | 0.093 | 34 | 0.141 |
| Random-sysname* | 33 | 0.537 | 36 | 0.046 | 34 | 0.018 | 32 | 0.462 |
| prismSrc* | 34 | 0.514 | 23 | 0.941 | 20 | 0.452 | 30 | 0.527 |
| *HuaweiTSC_EE_Metric* | – | – | 6 | 0.992 | – | – | – | – |
| *slide** | – | – | 13 | 0.982 | – | – | – | – |

Table 25: Correlations with WMT DA-SQM scores for all metrics on en→zh data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

| lang:<br>corr_fcn:<br>metric | avg-corr | | ja→en<br>pearson<br>task1 | | ja→en<br>pearson<br>task2 | | ja→en<br>acc-t<br>task3 | |
|---|---|---|---|---|---|---|---|---|
| *CometKiwi-XXL** | **1** | **0.798** | 2 | 0.984 | **1** | **0.474** | 2 | 0.578 |
| *CometKiwi-XL** | 2 | 0.795 | 3 | 0.982 | 4 | 0.446 | 6 | 0.573 |
| COMET | 3 | 0.787 | 16 | 0.968 | 5 | 0.445 | 3 | 0.576 |
| CometKiwi* | 4 | 0.787 | 10 | 0.975 | 3 | 0.455 | 10 | 0.568 |
| *cometoid22-wmt23** | 5 | 0.786 | 19 | 0.966 | 7 | 0.435 | 15 | 0.560 |
| KG-BERTScore* | 6 | 0.784 | 9 | 0.975 | 2 | 0.455 | 12 | 0.561 |
| *MetricX-23-QE-c** | 7 | 0.780 | 21 | 0.965 | 11 | 0.418 | 7 | 0.572 |
| BLEURT-20 | 8 | 0.778 | 22 | 0.964 | 6 | 0.436 | 9 | 0.570 |
| *MetricX-23-QE-b** | 9 | 0.777 | 12 | 0.972 | 16 | 0.383 | 4 | 0.575 |
| cometoid22-wmt22* | 10 | 0.776 | 25 | 0.946 | 8 | 0.432 | 19 | 0.550 |
| *MetricX-23-c* | 11 | 0.775 | 11 | 0.972 | 22 | 0.342 | 24 | 0.547 |
| *cometoid22-wmt21** | 12 | 0.774 | 26 | 0.944 | 9 | 0.431 | 20 | 0.549 |
| XCOMET-Ensemble | 13 | 0.774 | 24 | 0.947 | 12 | 0.410 | 5 | 0.574 |
| *MetricX-23-b* | 14 | 0.768 | 29 | 0.938 | 21 | 0.343 | **1** | **0.578** |
| MetricX-23-QE* | 15 | 0.768 | 30 | 0.936 | 20 | 0.344 | 11 | 0.567 |
| MS-COMET-QE-22* | 16 | 0.767 | 34 | 0.916 | 14 | 0.388 | 22 | 0.548 |
| XCOMET-QE-Ensemble* | 17 | 0.766 | 31 | 0.935 | 13 | 0.388 | 16 | 0.557 |
| MetricX-23 | 18 | 0.762 | 33 | 0.918 | 24 | 0.332 | 8 | 0.572 |
| YiSi-1 | 19 | 0.749 | 6 | 0.978 | 15 | 0.383 | 13 | 0.561 |
| *XCOMET-XL* | 20 | 0.748 | 32 | 0.922 | 25 | 0.327 | 23 | 0.547 |
| XLsim | 21 | 0.745 | **1** | **0.989** | 23 | 0.342 | 18 | 0.552 |
| *XCOMET-XXL* | 22 | 0.743 | 27 | 0.941 | 18 | 0.352 | 31 | 0.492 |
| GEMBA-MQM* | 23 | 0.739 | 4 | 0.982 | 10 | 0.421 | 34 | 0.395 |
| prismRef | 24 | 0.736 | 13 | 0.971 | 19 | 0.351 | 17 | 0.557 |
| mre-score-labse-regular | 25 | 0.734 | 5 | 0.980 | 33 | 0.186 | 21 | 0.548 |
| BERTscore | 26 | 0.732 | 7 | 0.977 | 17 | 0.357 | 14 | 0.560 |
| tokengram_F | 27 | 0.714 | 18 | 0.967 | 27 | 0.290 | 25 | 0.546 |
| chrF | 28 | 0.712 | 20 | 0.966 | 26 | 0.292 | 26 | 0.545 |
| f200spBLEU | 29 | 0.708 | 23 | 0.955 | 29 | 0.226 | 28 | 0.528 |
| embed_llama | 30 | 0.701 | 14 | 0.969 | 31 | 0.203 | 29 | 0.524 |
| eBLEU | 31 | 0.694 | 15 | 0.969 | 32 | 0.202 | 27 | 0.530 |
| BLEU | 32 | 0.660 | 28 | 0.939 | 30 | 0.221 | 30 | 0.517 |
| Random-sysname* | 33 | 0.537 | 36 | 0.288 | 35 | 0.061 | 32 | 0.481 |
| prismSrc* | 34 | 0.514 | 37 | -0.747 | 34 | 0.171 | 33 | 0.470 |
| *HuaweiTSC_EE_Metric* | – | – | 17 | 0.967 | – | – | – | – |
| *slide** | – | – | 8 | 0.976 | – | – | – | – |
| MaTESe | – | – | 35 | 0.904 | 28 | 0.242 | 35 | 0.326 |

Table 26: Correlations with WMT DA-SQM scores for all metrics on ja→en data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

| lang:<br>corr_fcn:<br>metric | avg-corr | | zh→en<br>pearson<br>task1 | | zh→en<br>pearson<br>task2 | | zh→en<br>acc-t<br>task3 | |
|---|---|---|---|---|---|---|---|---|
| *CometKiwi-XXL** | **1** | **0.798** | **1** | **0.938** | 3 | 0.435 | 4 | 0.540 |
| *CometKiwi-XL** | 2 | 0.795 | 3 | 0.936 | 6 | 0.427 | 9 | 0.535 |
| COMET | 3 | 0.787 | 21 | 0.811 | 11 | 0.378 | 15 | 0.525 |
| CometKiwi* | 4 | 0.787 | 4 | 0.931 | **1** | **0.460** | 10 | 0.534 |
| *cometoid22-wmt23** | 5 | 0.786 | 9 | 0.913 | 10 | 0.402 | 16 | 0.523 |
| KG-BERTScore* | 6 | 0.784 | 5 | 0.927 | 2 | 0.448 | 18 | 0.521 |
| *MetricX-23-QE-c** | 7 | 0.780 | 14 | 0.843 | 13 | 0.373 | 6 | 0.537 |
| BLEURT-20 | 8 | 0.778 | 25 | 0.766 | 17 | 0.331 | 21 | 0.520 |
| *MetricX-23-QE-b** | 9 | 0.777 | 17 | 0.823 | 21 | 0.298 | **1** | **0.544** |
| cometoid22-wmt22* | 10 | 0.776 | 8 | 0.918 | 5 | 0.432 | 19 | 0.520 |
| *MetricX-23-c* | 11 | 0.775 | 7 | 0.924 | 16 | 0.339 | 24 | 0.512 |
| *cometoid22-wmt21** | 12 | 0.774 | 10 | 0.908 | 7 | 0.419 | 20 | 0.520 |
| XCOMET-Ensemble | 13 | 0.774 | 20 | 0.816 | 18 | 0.322 | 7 | 0.537 |
| *MetricX-23-b* | 14 | 0.768 | 26 | 0.759 | 26 | 0.261 | 3 | 0.540 |
| MetricX-23-QE* | 15 | 0.768 | 24 | 0.770 | 22 | 0.284 | 5 | 0.538 |
| MS-COMET-QE-22* | 16 | 0.767 | 6 | 0.927 | 8 | 0.418 | 22 | 0.519 |
| XCOMET-QE-Ensemble* | 17 | 0.766 | 22 | 0.803 | 19 | 0.315 | 17 | 0.522 |
| MetricX-23 | 18 | 0.762 | 30 | 0.735 | 24 | 0.264 | 8 | 0.536 |
| YiSi-1 | 19 | 0.749 | 31 | 0.715 | 25 | 0.263 | 28 | 0.511 |
| *XCOMET-XL* | 20 | 0.748 | 28 | 0.758 | 27 | 0.254 | 27 | 0.512 |
| XLsim | 21 | 0.745 | 32 | 0.702 | 33 | 0.218 | 26 | 0.512 |
| *XCOMET-XXL* | 22 | 0.743 | 23 | 0.787 | 23 | 0.275 | 39 | 0.463 |
| GEMBA-MQM* | 23 | 0.739 | 11 | 0.873 | 14 | 0.370 | 41 | 0.356 |
| prismRef | 24 | 0.736 | 41 | 0.632 | 31 | 0.229 | 25 | 0.512 |
| mre-score-labse-regular | 25 | 0.734 | 18 | 0.817 | 38 | 0.146 | 30 | 0.509 |
| BERTscore | 26 | 0.732 | 33 | 0.702 | 30 | 0.236 | 23 | 0.515 |
| tokengram_F | 27 | 0.714 | 37 | 0.670 | 37 | 0.167 | 31 | 0.503 |
| chrF | 28 | 0.712 | 35 | 0.701 | 35 | 0.168 | 32 | 0.503 |
| f200spBLEU | 29 | 0.708 | 39 | 0.651 | 39 | 0.139 | 36 | 0.483 |
| embed_llama | 30 | 0.701 | 34 | 0.702 | 41 | 0.123 | 34 | 0.494 |
| eBLEU | 31 | 0.694 | 42 | 0.629 | 42 | 0.107 | 35 | 0.494 |
| BLEU | 32 | 0.660 | 43 | 0.610 | 40 | 0.134 | 37 | 0.475 |
| Random-sysname* | 33 | 0.537 | 44 | -0.144 | 43 | -0.026 | 40 | 0.446 |
| prismSrc* | 34 | 0.514 | 45 | -0.457 | 28 | 0.248 | 38 | 0.471 |
| *HuaweiTSC_EE_Metric* | – | – | 19 | 0.816 | – | – | – | – |
| *instructscore* | – | – | 38 | 0.652 | 32 | 0.227 | 42 | 0.342 |
| *slide** | – | – | 12 | 0.863 | – | – | – | – |
| Calibri-COMET22 | – | – | 27 | 0.759 | 20 | 0.313 | 13 | 0.529 |
| Calibri-COMET22-QE* | – | – | 13 | 0.854 | 12 | 0.375 | 11 | 0.530 |
| MEE4 | – | – | 40 | 0.632 | 36 | 0.168 | 33 | 0.498 |
| MaTESe | – | – | 29 | 0.739 | 34 | 0.201 | 43 | 0.319 |
| docWMT22CometDA | – | – | 15 | 0.836 | 15 | 0.345 | 12 | 0.530 |
| docWMT22CometKiwiDA* | – | – | 2 | 0.938 | 9 | 0.403 | 2 | 0.542 |
| mbr-metricx-qe* | – | – | 16 | 0.827 | 4 | 0.435 | 14 | 0.526 |
| sescoreX | – | – | 36 | 0.695 | 29 | 0.238 | 29 | 0.509 |

Table 27: Correlations with WMT DA-SQM scores for all metrics on zh→en data. Rows are sorted by the overall average correlation across all 25 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

## D Additional figures

Figures 9-14 show the (log) p-value of one-sided paired t-test on the MQM scores against the score difference of each metric for each system pair in each translation direction. Figures 15-20 show the (log) p-value of significance test with bootstrap resampling on the metric scores against the score difference of that metric for each system pair in each translation direction.

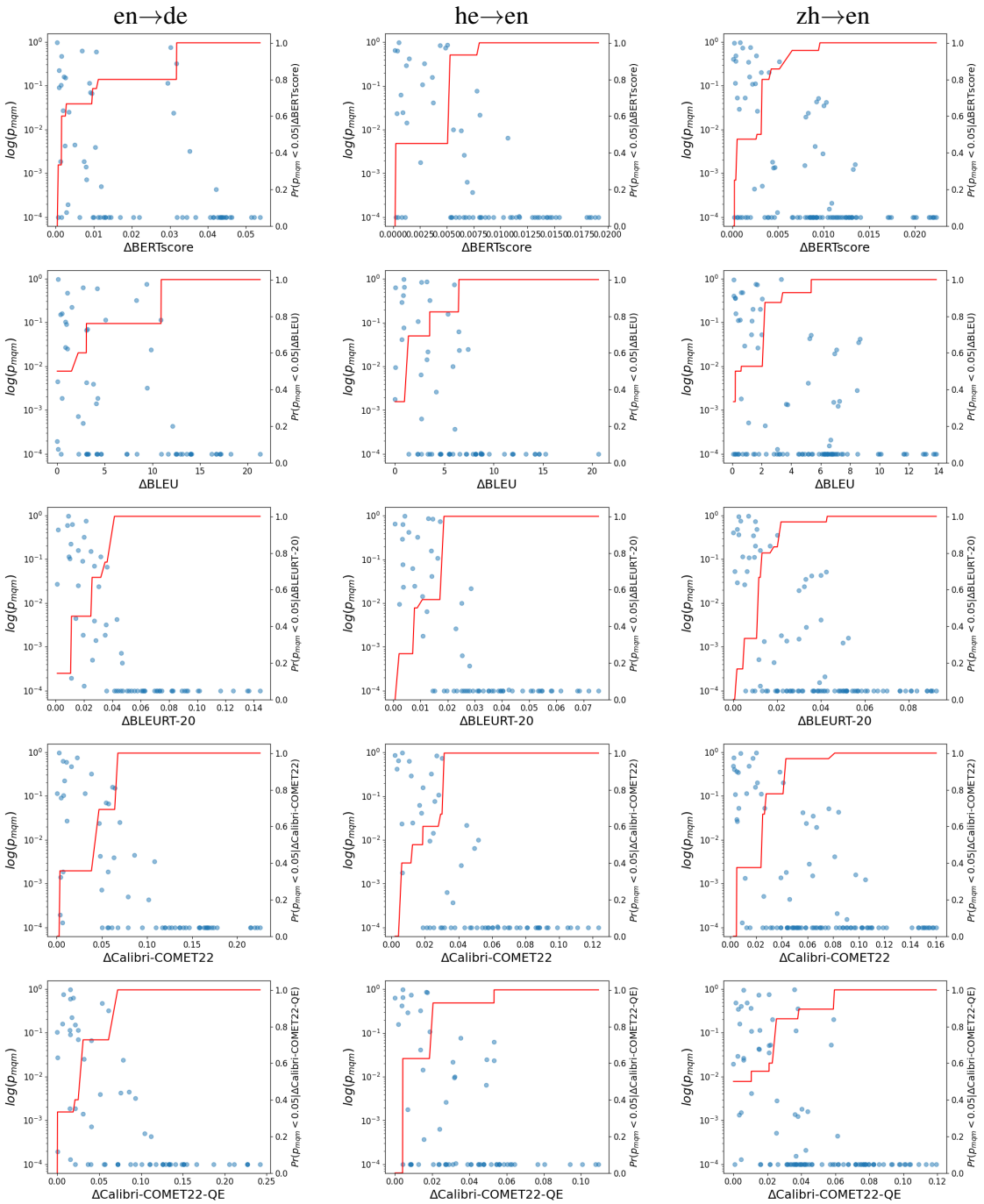


Figure 9: Log p-value of one-sided paired t-test on MQM scores ($p_{mqm}$) against the score difference of each metric (top to bottom: BERTScore, BLEU, BLEURT-20, CALIBRI-COMET22, CALIBRI-COMET22-QE) for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of $p_{mqm}$ are rounded up to 0.0001 when they are less than 0.0001.

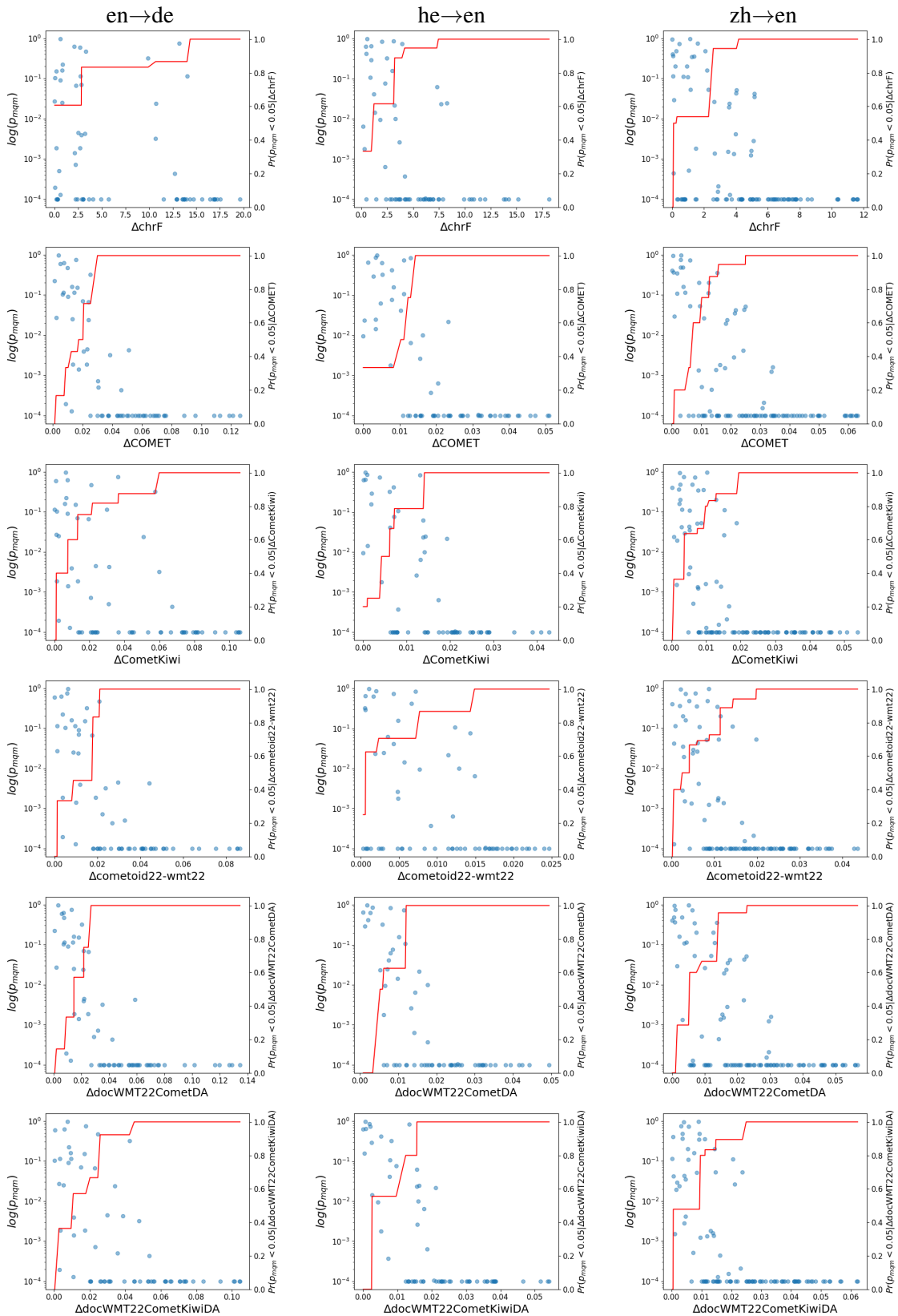


Figure 10: Log p-value of one-sided paired t-test on MQM scores ($p_{mqm}$) against the score difference of each metric (top to bottom: CHRF, COMET, COMETKIWI, COMETOID22-WMT22, DOCWMT22COMETDA, DOCWMT22COMETKIWIDA) for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of $p_{mqm}$ are rounded up to 0.0001 when they are less than 0.0001.

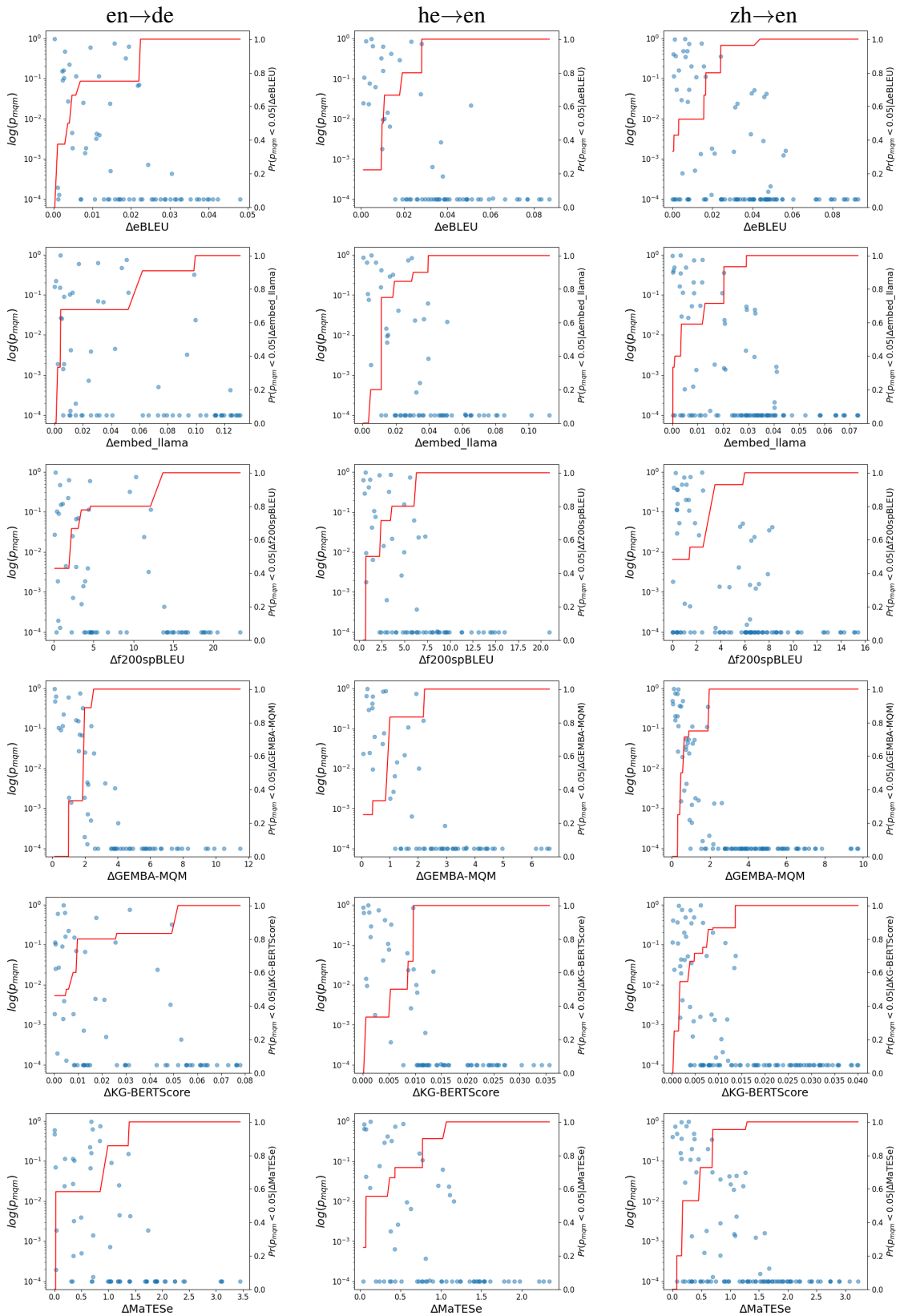


Figure 11: Log p-value of one-sided paired t-test on MQM scores ($p_{mqm}$) against the score difference of each metric (top to bottom: EBLEU, EMBED_LLAMA, F200SPBLEU, GEMBA-MQM, KG-BERTSCORE, MATESE) for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of $p_{mqm}$ are rounded up to 0.0001 when they are less than 0.0001.

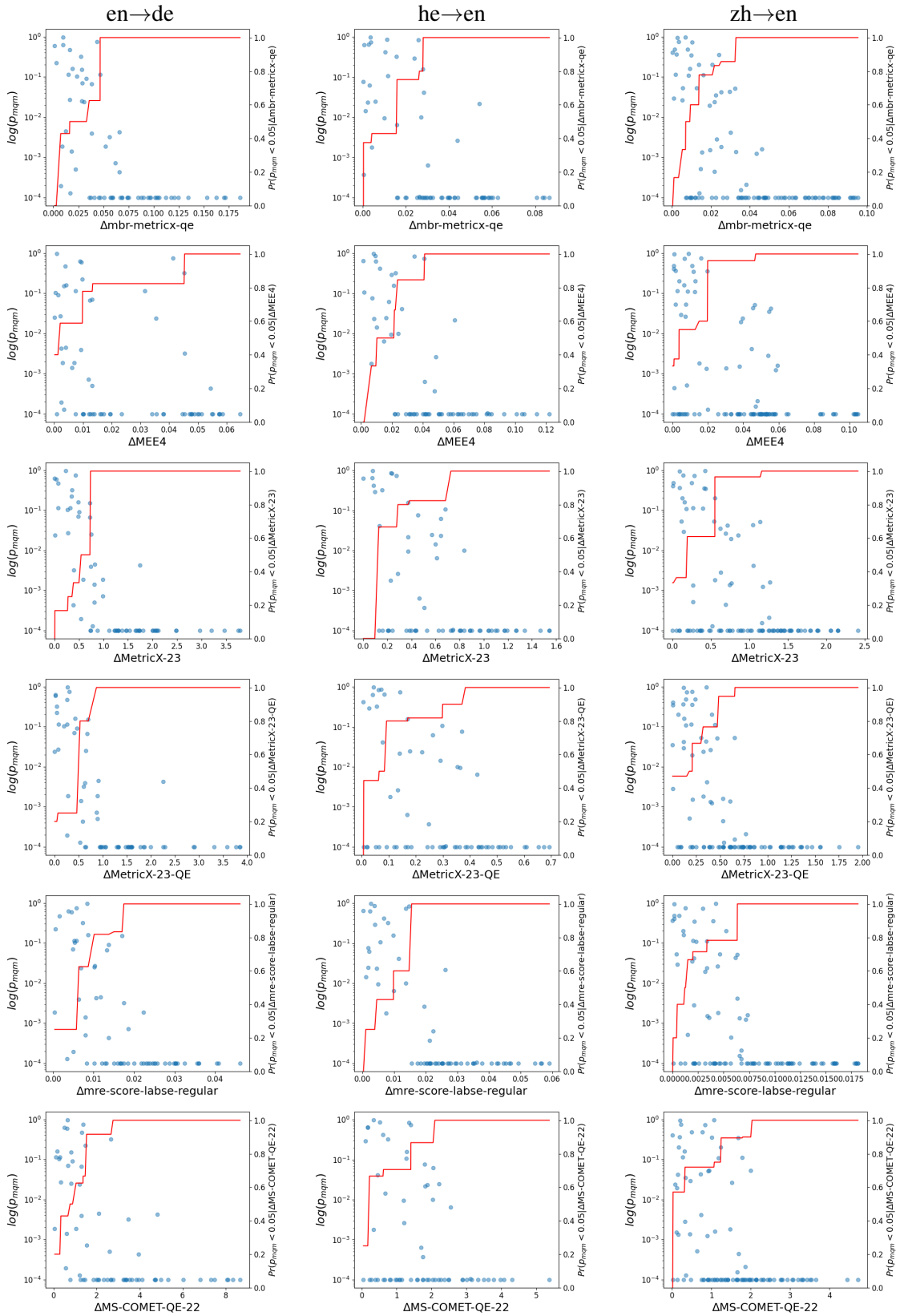


Figure 12: Log p-value of one-sided paired t-test on MQM scores ($p_{mqm}$) against the score difference of each metric (top to bottom: MBR-METRICX-QE, MEE4, METRICX-23, METRICX-23-QE, MRE-SCORE-LABSE-REGULAR, MS-COMET-QE-22) for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of $p_{mqm}$ are rounded up to 0.0001 when they are less than 0.0001.

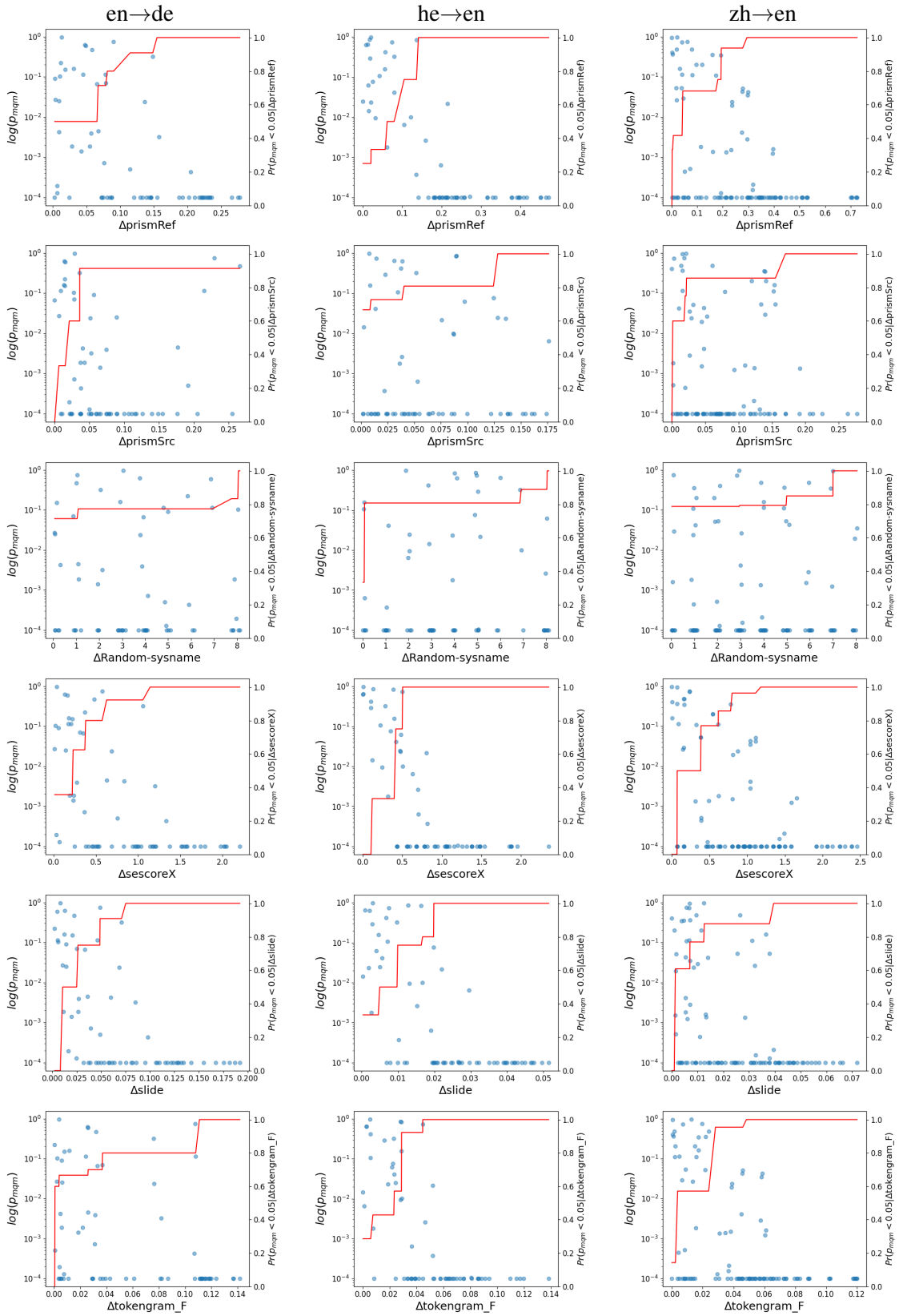


Figure 13: Log p-value of one-sided paired t-test on MQM scores ($p_{mqm}$) against the score difference of each metric (top to bottom: PRISMREF, PRISMSRC, RANDOM-SYSNAME, SESCOREX, SLIDE, TOKENGRAM_F) for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of $p_{mqm}$ are rounded up to 0.0001 when they are less than 0.0001.

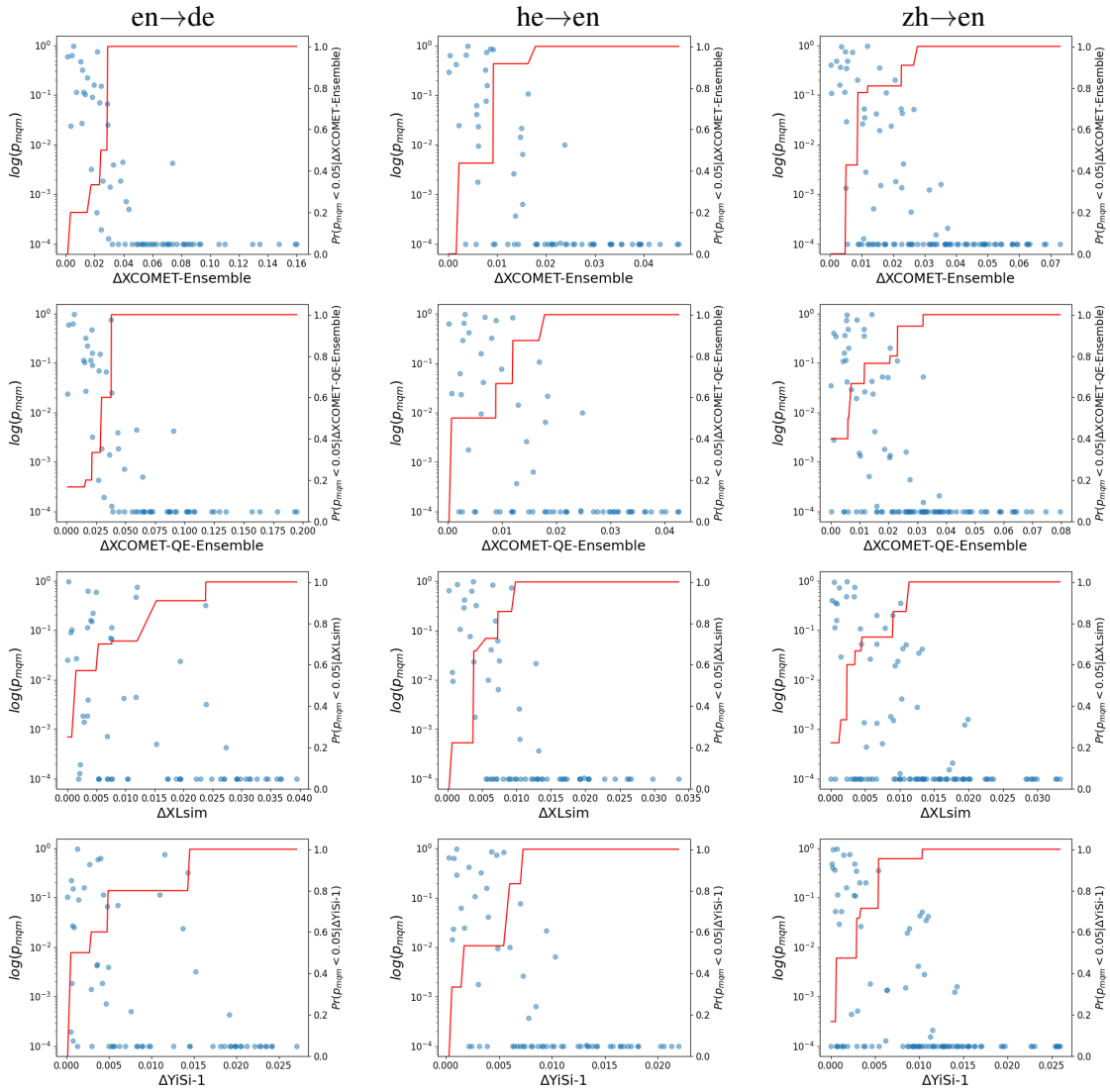


Figure 14: Log p-value of one-sided paired t-test on MQM scores ($p_{mqm}$) against the score difference of each metric (top to bottom: XCOMET-ENSEMBLE, XCOMET-QE-ENSEMBLE, XLSIM, YISI-1) for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of $p_{mqm}$ are rounded up to 0.0001 when they are less than 0.0001.

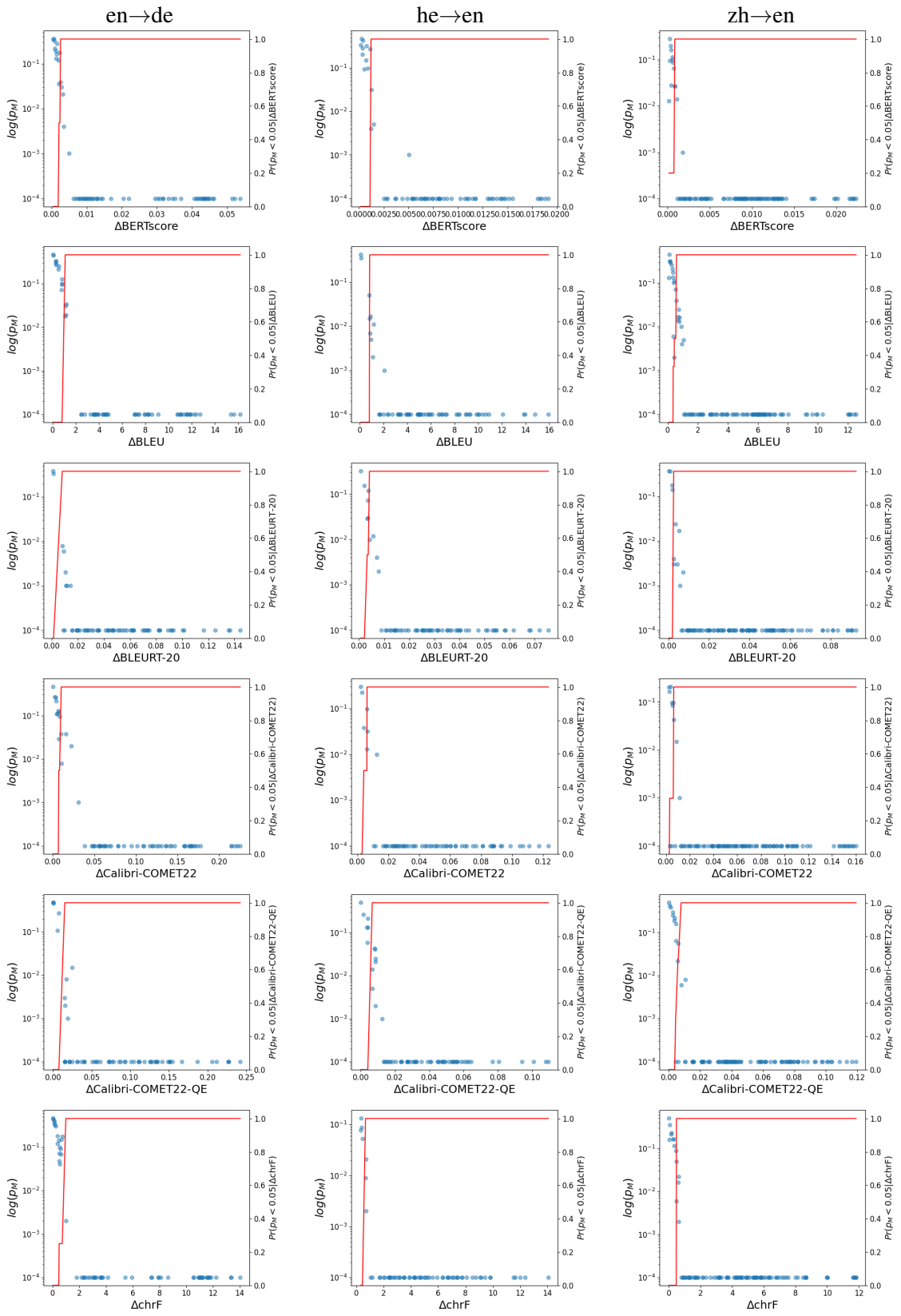


Figure 15: Log p-value of significance test with bootstrap resampling ($p_M$) on system-level metric scores against each metric (top to bottom: BERTSCORE, BLEU, BLEURT-20, CALIBRI-COMET22, CALIBRI-COMET22-QE, CHRF) score difference for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of $p_M$ are rounded up to 0.0001 when they are less than 0.0001.

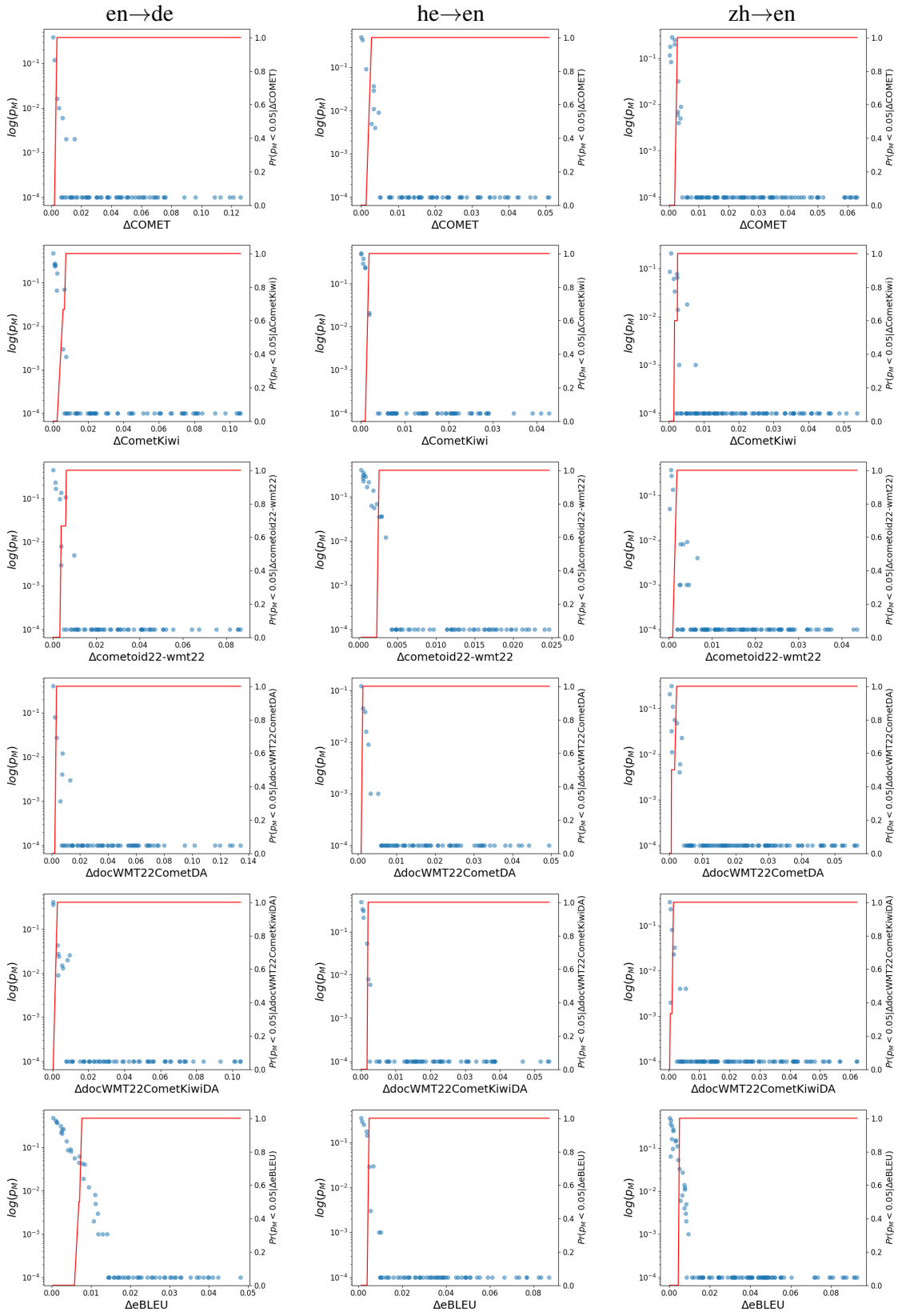


Figure 16: Log p-value of significance test with bootstrap resampling ($p_M$) on system-level metric scores against each metric (top to bottom: COMET, COMETKIWI, COMETOID22-WMT22, DOCWMT22COMETDA, DOCWMT22COMETKIWIDA, EBLEU) score difference for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of $p_M$ are rounded up to 0.0001 when they are less than 0.0001.

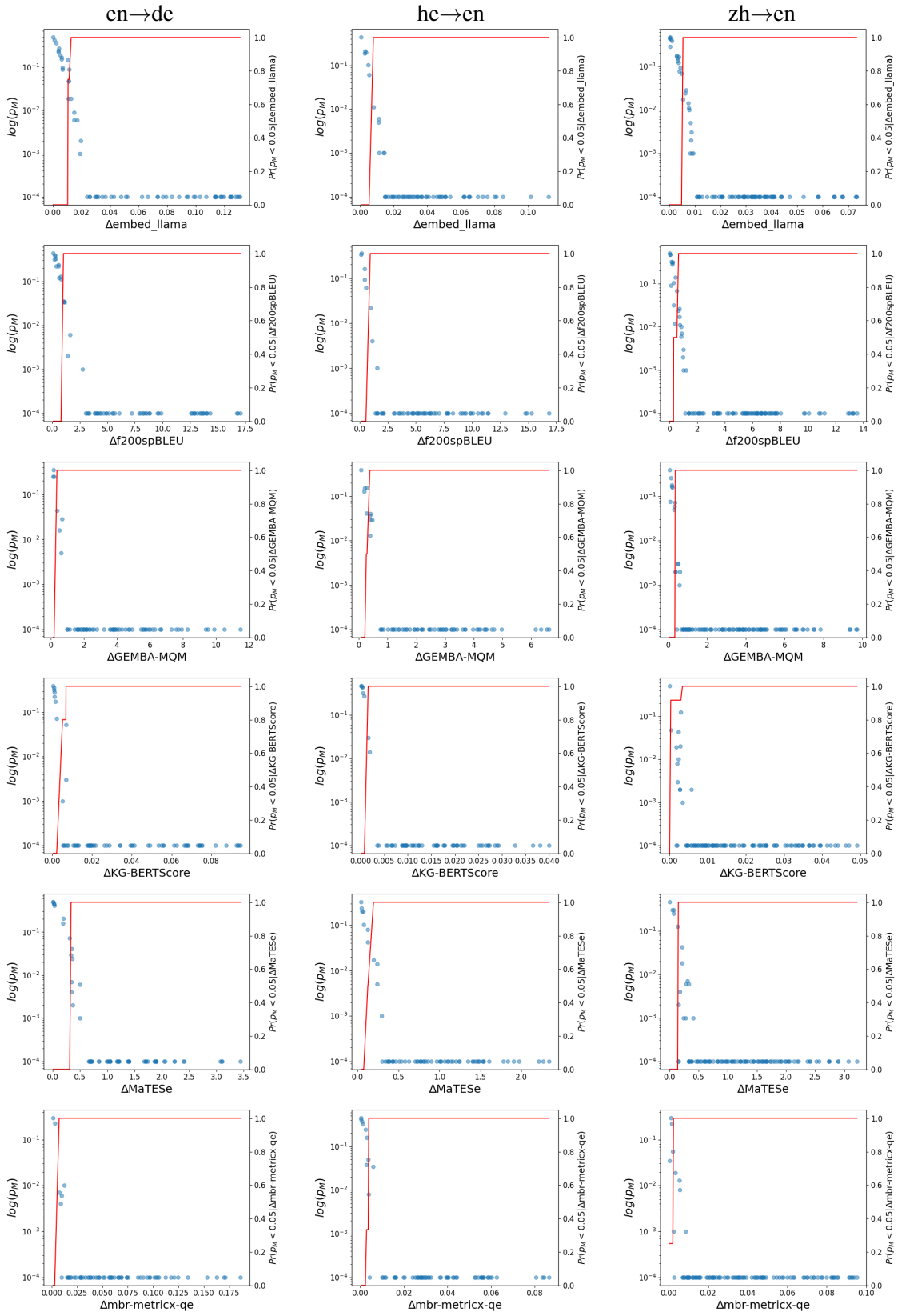


Figure 17: Log p-value of significance test with bootstrap resampling ($p_M$) on system-level metric scores against each metric (top to bottom: EMBED_LLAMA, F200SPBLEU, GEMBA-MQM, KG-BERTSCORE, MATESE, MBR-METRICX-QE) score difference for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of $p_M$ are rounded up to 0.0001 when they are less than 0.0001.

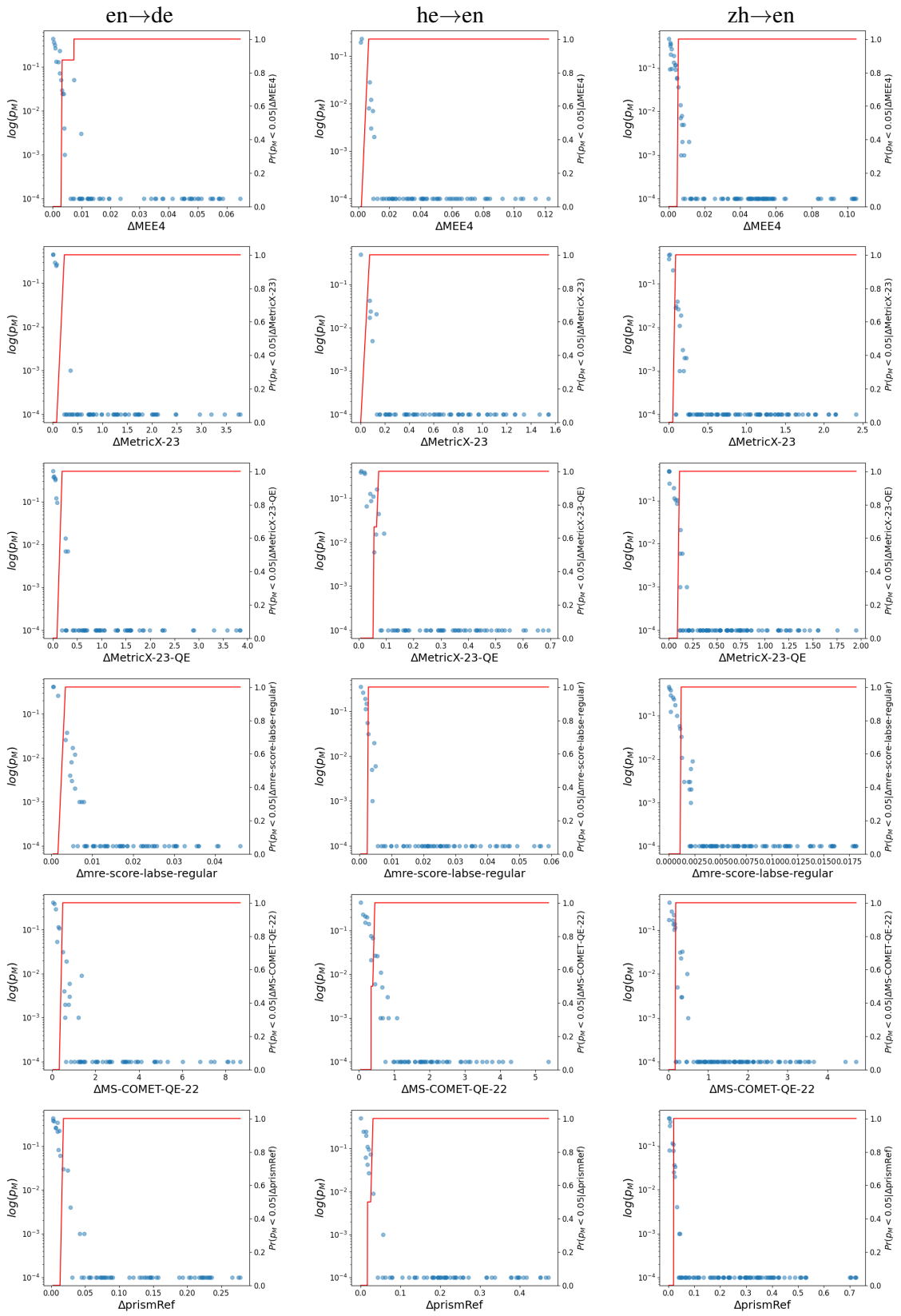


Figure 18: Log p-value of significance test with bootstrap resampling ($p_M$) on system-level metric scores against each metric (top to bottom: MEE4, METRICX-23, METRICX-23-QE, MRE-SCORE-LABSE-REGULAR, MS-COMET-QE-22, PRISMREF) score difference for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of $p_M$ are rounded up to 0.0001 when they are less than 0.0001.

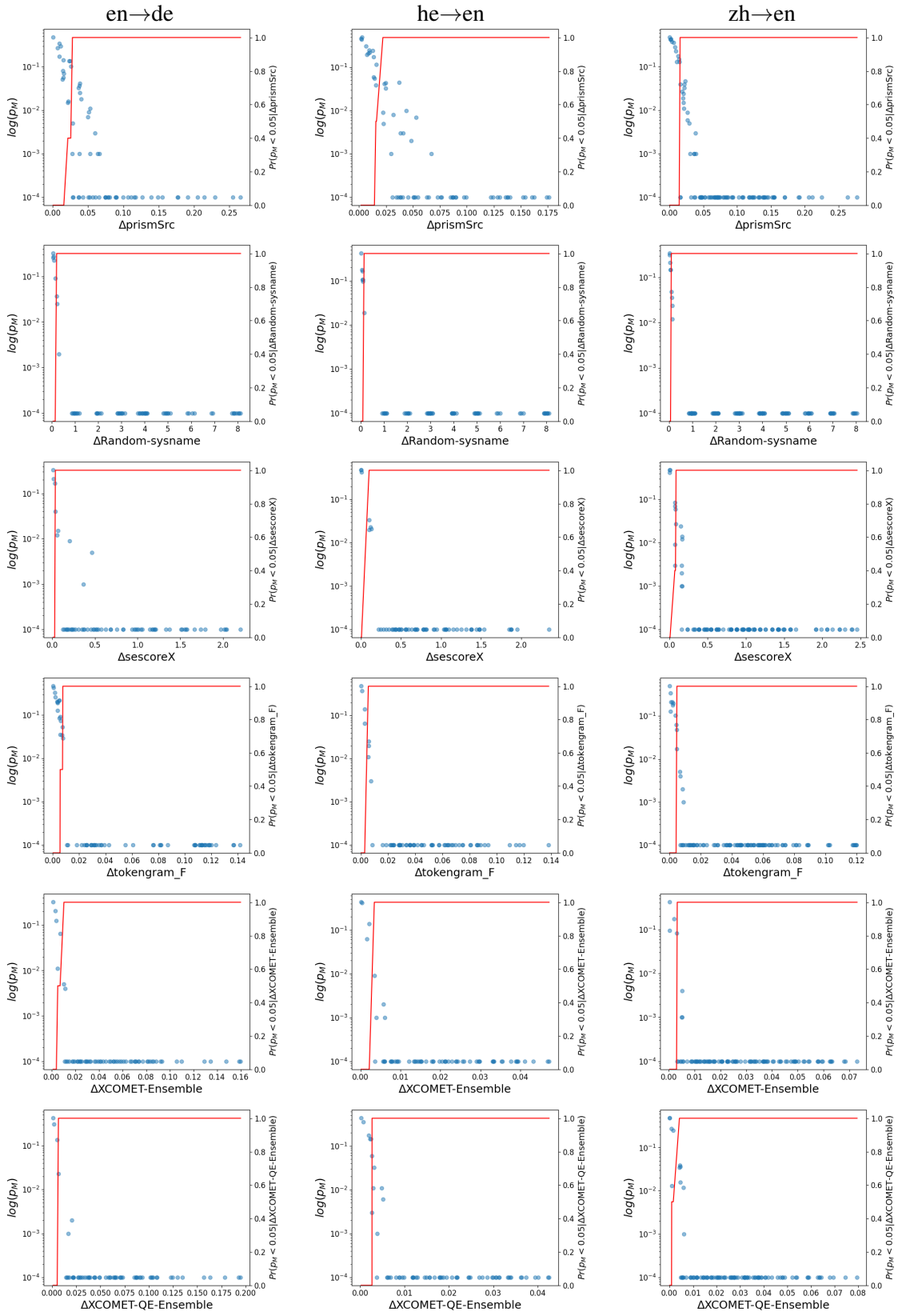


Figure 19: Log p-value of significance test with bootstrap resampling ($p_M$) on system-level metric scores against each metric (top to bottom: PRISMSRC, RANDOM-SYSNAME, SESCOREX, TOKENGRAM_F, XCOMET-ENSEMBLE, XCOMET-QE-ENSEMBLE) score difference for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of $p_M$ are rounded up to 0.0001 when they are less than 0.0001.

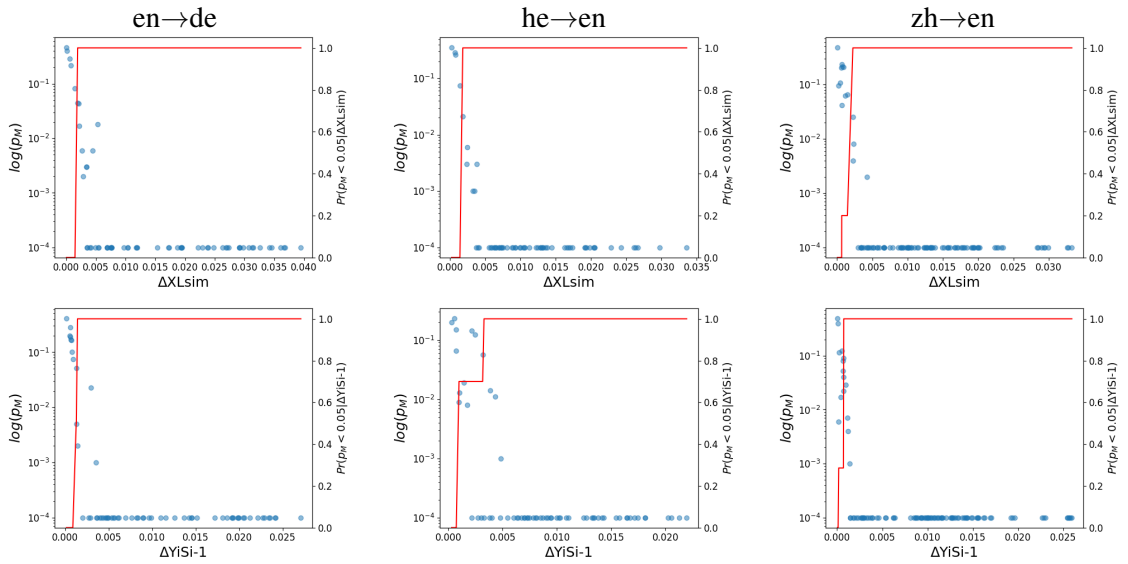


Figure 20: Log p-value of significance test with bootstrap resampling ($p_M$) on system-level metric scores against each metric (top to bottom: XLSIM, YISI-1) score difference for each system pair in each translation direction (left to right: en→de, he→en, zh→en). The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of $p_M$ are rounded up to 0.0001 when they are less than 0.0001.