Are LLMs Breaking MT Metrics?
Results of the WMT24 Metrics Shared Task

Markus Freitag[(1)], Nitika Mathur[(2)], Daniel Deutsch[(1)], Chi-kiu Lo 羅致翹[(3)], Eleftherios Avramidis[(4)], Ricardo Rei[(5)], Brian Thompson[(6)], Frédéric Blain[(7)], Tom Kocmi[(8)], Jiayi Wang[(9)], David I. Adelani[(10,11)], Marianna Buchicchio[(5)], Chrysoula Zerva[(12,13)], Alon Lavie[(14)]

[(1)]Google Research [(2)]Oracle [(3)]National Research Council Canada
[(4)]German Research Center for Artificial Intelligence (DFKI) [(5)]Unbabel [(6)]Amazon [(7)]Tilburg University
[(8)]Microsoft [(9)]University College London [(10)]McGill University [(11)]Mila - Quebec AI Institute
[(12)]Instituto Superior Técnico [(13)]Instituto de Telecomunicações [(14)]Phrase
wmt-metrics@googlegroups.com

Abstract

The WMT24 Metrics Shared Task evaluated the performance of automatic metrics for machine translation (MT), with a major focus on LLM-based translations that were generated as part of the WMT24 General MT Shared Task. As LLMs become increasingly popular in MT, it is crucial to determine whether existing evaluation metrics can accurately assess the output of these systems.

To provide a robust benchmark for this evaluation, human assessments were collected using Multidimensional Quality Metrics (MQM), continuing the practice from recent years. Furthermore, building on the success of the previous year, a challenge set subtask was included, requiring participants to design contrastive test suites that specifically target a metric's ability to identify and penalize different types of translation errors.

Finally, the meta-evaluation procedure was refined to better reflect real-world usage of MT metrics, focusing on pairwise accuracy at both the system- and segment-levels.

We present an extensive analysis on how well metrics perform on three language pairs: English→Spanish (Latin America), Japanese→Chinese, and English→German. The results strongly confirm the results reported last year, that fine-tuned neural metrics continue to perform well, even when used to evaluate LLM-based translation systems.

metric		avg corr
MetaMetrics-MT	**1**	**0.725**
MetricX-24-Hybrid	**1**	**0.721**
XCOMET	**1**	**0.719**
MetricX-24-Hybrid-QE*	2	0.714
gemba_esa*	2	0.711
XCOMET-QE*	3	0.695
COMET-22	3	0.688
BLEURT-20	3	0.686
MetaMetrics-MT-QE*	3	0.684
bright-qe*	4	0.681
BLCOM_1	4	0.664
sentinel-cand-mqm*	5	0.650
PrismRefMedium	5	0.646
PrismRefSmall	5	0.642
CometKiwi*	5	0.640
damonmonli	5	0.635
YiSi-1	6	0.630
BERTScore	7	0.617
MEE4	7	0.609
chrF	8	0.608
chrfS	8	0.606
spBLEU	9	0.593
BLEU	9	0.589
XLsimMqm*	10	0.515
sentinel-src-mqm*	10	0.513
sentinel-ref-mqm	10	0.513

Table 1: Official ranking of primary submissions to the WMT24 Metric Task. The final score is the weighted average correlation over 6 different tasks. Starred metrics are reference-free, and underlined metrics are baselines. See Table 14 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 evaluate 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 ratings when scoring MT systems and human translations at the system and sentence level.

[1]https://www2.statmt.org/wmt24/metrics-task.html

Proceedings of the Ninth Conference on Machine Translation, pages 47–81
November 15-16, 2024 ©2024 Association for Computational Linguistics

The final ranking of this year's submitted primary metrics is shown in Table 1. Below are some of the key details and changes implemented for this year's Metrics Shared Task:

- **Language Pairs**: For this year, we focus on three language pairs, all on the paragraph-level: (i) English→German (en→de), English→Spanish (Latin America) (en→es), and Japanese→Chinese (ja→zh).

- **Human Evaluation**: Like last year, we collected our own human quality ratings for our three language pairs leveraging professional translators performing MQM annotations (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, we designed the meta-evaluation to evaluate metrics on how they are used in practice, by focusing on pairwise accuracy at the system- and segment-levels and removing Pearson correlation. At the system-level, we use a new statistic called soft pairwise accuracy (Thompson et al., 2024), and, like last year, we use pairwise accuracy with tie calibration (Deutsch et al., 2023) at the segment-level.

- **Challenge Sets Subtask**: The submission format of the challenge sets changed to provide for more flexibility on how the participants could challenge the metrics. In contrast to previous years, when the challenge items were evaluated in a rigid pairwise manner on whether the metric scores can distinguish between a good and a bad translation, this year's participants could provide single translations and then employ an evaluation concept of their own. This year's subtask features 4 submissions that test the ability of the metrics to evaluate MT outputs on African languages, the biomedical domain, on more than a hundred linguistically-motivated phenomena, as well as on low- to mid-quality outputs and specific challenges (empty strings, wrong/mixed language output and language variants).

- **Understand Magnitude of Score Difference**: Similar to last 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:

- Two metametrics (which are both ensemble metrics), MetricX-24-Hybrid and XCOMET, are the winners of the WMT24 Metrics Shared Task (Table 1);

- Fine-tuned neural metrics continue to be strong in performance and are effective quality estimators, even for LLM-based translations;

- Results from the challenge sets independently suggest that it is important for metric researchers to test the performance of metrics in diverse collections of linguistic phenomena, languages and domains, including low-resource languages, mixed languages and irregular outputs, and on a wide range of translation quality, in order to minimize anomalous and unexpected behaviours of metrics (Section 9).

The rest of the paper is organized as follows: Section 2 describes the test data. 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 summarizes our results for the WMT24 General MT Shared Task language-pairs based on their new ESA human evaluation methodology (Kocmi et al., 2024c). Section 9 presents a description of the submitted challenge sets along with their findings. Finally, Section 10 summarizes our most important conclusions.

[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

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 WMT24 General MT Shared Task (Kocmi et al., 2024a). The domains cover news, literary, speech, and social. We do not provide any sentence splitting, thus many segments contain multiple sentences. Each language pair contains a comparable number of sentences from each domain, resulting in reasonably balanced test sets. Data statistics can be seen in Table 2. The language pairs en→de and en→es have the same source segments; ja→zh consists of segments from only 3 different domains.

	news	literary	speech	social
#tokens				
en→{de,es}	9,268	9,601	9,611	9,829
ja→zh	14,896	14,541	11,025	
#docs (#segments/doc)				
en→{de,es}	17 (8.8)	8 (25.8)	111 (1.0)	34 (15.6)
ja→zh	45 (6.0)	15 (21.1)	136 (1.0)	
#sents (#sents/doc)				
en→{de,es}	333 (19.6)	607 (75.9)	685 (6.2)	759 (22.3)
ja→zh	634 (14.1)	875 (58.3)	332 (2.4)	

Table 2: Test set statistics split by domain. Statistics are calculated on the source side.

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

For more details regarding the test sets, we refer the reader to the WMT24 General MT Shared Task findings paper (Kocmi et al., 2024a). All data has been released and can be downloaded[5].

3 MQM Human Evaluation

Automatic metrics are commonly evaluated by measuring correlations with corresponding human ratings. The quality of these human ratings is critical, and recent findings (Freitag et al., 2021) have shown that crowdsourced human ratings are not sufficiently reliable for evaluating high quality MT outputs. Furthermore, an evaluation schema based on MQM (Lommel et al., 2014), which requires explicit error annotation is more effective than 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 translation system submissions and language pairs which we believe are most interesting for evaluating current metrics. Instead of evaluating all MT system submissions, we restrict our human evaluation to the top scoring submissions, as determined based on baseline automatic scores. 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, Japanese→Chinese) or trusted and trained raters (English→Spanish). 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 & Japanese→Chinese

Annotations for en→de and ja→zh were sponsored and executed by Google, using 18 professional translators (10 for en→de, 8 for ja→zh) 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 3.

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

Table 3: Google's MQM error weighting.

Recent research demonstrated that rater assignment is crucial for reliable human evaluation and we adopted the suggested Pseudo-Side-by-Side

[5] https://github.com/wmt-conference/wmt24-news-systems

(pSxS) rater assignment as suggested in (Riley et al., 2024). 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 three metrics tasks.

3.2 English→Spanish (Latin America)

The annotations for the en→es (Latin America)[6] 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[7].

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 3.

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 4, you can see the number of segments and documents for each language pair and test set that we used for human evaluation. 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 5. It's important to note a non-intentional, but important difference in our human evaluation setting for the speech domain between the three language pairs. For English→German and English→Spanish, we asked human annotators to compare translations against the ASR output, which inadvertently disadvantaged participants who used audio input, including those providing human translations, as these translations rely on an error-free input. This is evident in the higher MQM scores for the speech domain for both language pairs for human translations and the dubformer system (which also utilizes audio input). However, for Japanese→Chinese, the human annotators compared against the cleaned human transcription. This mismatch was not intentional and we will discuss the impact on the correlation numbers in Section 6.

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 6.

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 sentence-BLEU values using the `sentence_bleu` Python function and for corpus-level BLEU we used `corpus_bleu` (both with default arguments[8]).

- **SPBLEU (NLLB Team et al., 2022)** are BLEU scores computed with subword tokenization by the standardized FLORES-200 Sentencepiece models. We used the command line SacreBLEU to compute the sentence level SPBLEU[9] and we averaged 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[10]) for segment-level scores and we average those scores to obtain a corpus-level score.

[6]Since the testset is for Spanish from Mexico rather than Spanish from Spain, the conducted annotations were collected taking that variant in consideration.

[7]see Unbabel Annotation Guidelines - Typology 3.0

[8]nrefs.1|case.mixed|lang.LANGPAIR|tok.13a|smooth.exp| version.2.3.0. For to-zh and to-ja language pairs, we use tok.zh and tok.ja-mecab

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

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

language	news	social	speech	literary
en→de	90/149 (17/17)	258/531 (34/34)	111/111 (1/1)	27/206 (8/8)
en→es	124/149 (14/17)	281/531 (20/34)	107/111 (1/1)	110/206 (5/8)
ja→zh	255/269 (45/45)	n/a	136/136 (1/1)	168/316 (15/15)

Table 4: Numbers of MQM-annotated segments per domain (number of docs in brackets).

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-22 (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 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 the WMT22 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.

PRISMREFSMALL AND PRISMREFMEDIUM (Thompson and Post, 2020a,b) are both reference-based PRISM that uses a multilingual MT model in zero-shot paraphrase model to score the candidate translation conditioned on the reference, and the reference conditioned on the candidate translation, and averages the two scores. As LLMs have become quite capable multi-lingual MT models, we opted to use Llama3.1 (Llama Team, 2024) as the underlying MT model this year. PRISMREFSMALL corresponds to Llama3.1 8B and PRISMREFMEDIUM corresponds to Llama3.1 70B. The long context window of LLMs allows us to compute scores for entire documents, while still averaging scores for each sentence to produce sentence-level scores (Vernikos et al., 2022). We chunked longer documents into sub-documents of up to 10 sentences, and added a penalty for producing no output.

YISI-1 (Lo, 2019) is an 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.

BLCOM_1 and BLCOM Unfortunately, we have no information about these submission.

BRIGHT-QE is a referenceless metric, which uses the XLM-XL encoder to perform multi-stage fine-tuning according to the XCOMET framework. In the first stage of training, we used DA 2017 2022 corpus, and gradually reduced the weight of REF-based loss with the idea of curriculum learning, trying to reduce the model's dependence on reference and better align the semantics of the translation and source text; in the second stage, we used batch softmax to normalize scores, and introduced KL divergence loss to learn to modify the minor rank error that MSE loss cannot solve, so as to obtain better Pearson correlation; finally, we further fine-tuned on high-quality MQM corpus to achieve better consistency with human expert MQM.

System	English→German ↓ all	news	social	speech	literary
Dubformer	1.58	1.29	0.60	4.22	1.15
GPT-4	1.58	1.39	0.88	3.60	0.69
Unbabel-Tower70B	1.65	1.99	0.78	3.46	1.41
ONLINE-B	1.81	1.48	1.22	3.59	1.30
TranssionMT	1.81	1.24	1.18	3.87	1.33
refB	1.84	1.38	0.80	4.92	0.81
Mistral-Large	1.93	1.95	1.12	3.91	1.46
CommandR-plus	2.01	2.40	1.07	3.95	1.74
refA	2.12	1.84	1.01	4.96	2.04
Gemini-1.5-Pro	2.20	1.29	1.93	2.90	4.97
ONLINE-W	2.22	1.32	1.75	4.09	2.12
Claude-3.5	2.28	1.00	1.23	6.04	1.13
IOL_Research	2.39	1.66	1.61	4.91	2.01
Aya23	3.09	2.69	2.20	5.71	2.26
ONLINE-A	3.30	1.93	2.29	6.88	2.85
Llama3-70B	3.62	2.91	2.28	7.08	4.76
IKUN	3.86	4.35	2.36	7.09	3.48
IKUN-C	5.07	3.39	3.34	9.87	7.63
MSLC	13.46	11.54	8.24	26.80	15.29

System	English→Spanish ↓ all	news	social	speech	literary
GPT-4	0.12	0.03	0.14	0.24	0.03
Unbabel-Tower70B	0.20	0.21	0.04	0.68	0.14
Claude-3.5	0.26	0.06	0.21	0.60	0.29
Mistral-Large	0.26	0.16	0.28	0.50	0.12
Gemini-1.5-Pro	0.39	0.18	0.56	0.54	0.06
Dubformer	0.43	0.29	0.07	2.00	0.01
Llama3-70B	0.52	0.10	0.28	2.17	0.02
refA	0.55	0.20	0.12	2.42	0.20
IOL_Research	0.57	0.44	0.33	1.39	0.56
CommandR-plus	0.62	0.50	0.34	0.52	1.55
ONLINE-W	0.64	0.17	0.27	2.36	0.46
IKUN	0.94	0.86	0.74	1.01	1.46
ONLINE-B	1.08	1.01	0.59	1.76	1.77
Aya23	1.52	1.52	1.09	2.03	2.12
MSLC	6.80	4.09	4.63	10.99	11.36

System	Japanese→Chinese ↓ all	news	speech	literary
Claude-3.5	1.22	0.76	2.96	0.76
refA	1.32	0.77	3.15	0.77
GPT-4	1.45	0.82	3.25	0.82
DLUT_GTCOM	1.52	1.06	3.66	1.06
Unbabel-Tower70B	1.69	1.16	3.53	1.16
Gemini-1.5-Pro	1.78	0.84	3.80	0.84
CommandR-plus	1.91	1.28	4.61	1.28
IOL_Research	2.10	1.14	4.82	1.14
Aya23	3.03	1.86	6.44	1.86
Llama3-70B	3.07	2.16	6.16	2.16
Team-J	3.91	2.02	8.46	2.02
NTTSU	4.34	2.11	10.51	2.11
ONLINE-B	5.27	3.72	9.52	3.72
IKUN-C	6.60	3.45	14.41	3.45
MSLC	9.19	4.01	19.04	4.01

Table 5: MQM human evaluations for generalMT2024. Lower average error counts represent higher MT quality. Systems above any solid line are significantly better than those below, based on all domains with $p < 0.05$.

CHRFS (Mukherjee and Shrivastava, 2024) is an unsupervised reference-based metric, a semantic version of CHRF++ that integrates sentence embeddings to evaluate translation quality more comprehensively. By combining traditional character and word n-gram analysis with semantic information derived from embeddings, CHRFS captures both syntactic accuracy and sentence-level semantics.

DAMONMONLI and MONMONLI is a proof-of-concept of multiple ideas. A multi-lingual NLI model is used to extract embeddings for (mt, src) and (mt, ref) pairs, based on findings of Chen and Eger (2023). A multi-task learning approach is employed where different human annotations from WMT22 and WMT23 are used as different tasks. For each task, it uses a separate regression head that learns a monotonic function of the metric's score(Runje and Shankaranarayana, 2023). The main metric "DAMONMONLI" also includes a domain adversarial loss (Ganin and Lempitsky, 2015) to make metric representations robust against shifts in MT systems and language pairs.

GEMBA-ESA (Kocmi and Federmann, 2023) is an extension of previous work on an LLM-based metric, with an updated prompt to reflect the new human evaluation protocol ESA (Kocmi et al., 2024c) used at WMT General MT task. It contains a two-step approach where in the first step, MQM error spans are collected and in a second step, the final score is assigned.

MEE4 (Mukherjee and Shrivastava, 2023a) 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.

METAMETRICS-MT (Anugraha et al., 2024; Winata et al., 2024) is a machine translation

	metric	broad category	supervised	ref. free	citation	availability (https://github.com/)
baselines	BLEU	lexical overlap			Papineni et al. (2002)	mjpost/sacrebleu
	spBleu	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-20	fine-tuned metric	✓		Sellam et al. (2020)	google-research/bleurt
	Comet-22	fine-tuned metric	✓		Rei et al. (2022a)	Unbabel/COMET
	CometKiwi	fine-tuned metric	✓	✓	Rei et al. (2022b)	Unbabel/COMET
	prismRefSmall	MT-model metric			Thompson and Post (2020a,b)	thompsonb/prism
	prismRefMedium	MT-model metric			Thompson and Post (2020a,b)	thompsonb/prism
	YiSi-1	embedding similarity			Lo (2019)	chikiulo/yisi
primary submissions	BLCOM_1	N/A	N/A	N/A	N/A	(not available)
	BRIGHT-QE	fine-tuned metric	✓	✓	N/A	https://bright.pcl.ac.cn/en/
	chrfS	lexical and embedding similarity			(Mukherjee and Shrivastava, 2024)	AnanyaCoder/chrF-S
	CometKiwi-XXL	fine-tuned metric	✓	✓	Rei et al. (2023)	Unbabel/COMET
	damonmonli	finetuned metric	✓		N/A	(not available)
	GEMBA-ESA	LLM prompt-based metric		✓	Kocmi and Federmann (2023)	MicrosoftTranslator/GEMBA
	MEE4	lexical & embedding similarity			Mukherjee and Shrivastava (2023b)	AnanyaCoder/WMT22Submission
	MetaMetrics-MT	ensemble metric	✓		Anugraha et al. (2024)	meta-metrics/metametrics
	MetaMetrics-MT-QE	ensemble metric	✓	✓	Anugraha et al. (2024)	gentaiscool/meta-metrics
	MetricX-24-Hybrid	fine-tuned metric	✓		Juraska et al. (2024)	google-research/metricx
	MetricX-24-Hybrid-QE	fine-tuned metric	✓	✓	Juraska et al. (2024)	google-research/metricx
	sentinel-cand-mqm	fine-tuned metric	✓	✓	Perrella et al. (2024)	SapienzaNLP/guardians-mt-eval
	sentinel-ref-mqm	fine-tuned metric	✓		Perrella et al. (2024)	SapienzaNLP/guardians-mt-eval
	sentinel-src-mqm	fine-tuned metric	✓	✓	Perrella et al. (2024)	SapienzaNLP/guardians-mt-eval
	XCOMET	fine-tuned metric	✓		Guerreiro et al. (2023)	Unbabel/COMET
	XCOMET-QE	fine-tuned metric	✓	✓	Guerreiro et al. (2023)	Unbabel/COMET
	XLsimMQM	fine-tuned metric	✓		Mukherjee and Shrivastava (2023b)	AnanyaCoder/XLsim

Table 6: Baseline metrics and primary submissions for the metrics task. Supervised metrics are trained on MT evaluation data such as DA or MQM scores.

(MT) metric developed from our METAMETRICS (Winata et al., 2024), specifically designed to better align with human preferences using Bayesian optimization with Gaussian Processes (GP). By systematically integrating multiple existing metrics, we create a sparse allocation that only includes metrics enhancing the overall correlation score. We optimize this metric by maximizing Kendall scores from the WMT shared task (MQM) 2020-2022. METAMETRICS-MT achieves state-of-the-art performance for reference-based metrics, while its reference-free variant, METAMETRICS-MT-QE, demonstrates competitive correlation with human scores in the WMT24 metric shared task. By strategically assigning weights to combined metrics, METAMETRICS-MT aims to be as competitive as, if not superior to, any individual metric. To address missing values when reference data is unavailable, we propose a hybrid variant, METAMETRICS-MT-HYBRID, which utilizes both metrics to compensate for the absence of reference data in the reference-based setting.

METRICX-24 (Juraska et al., 2024) is a learned regression-based metric that builds on top of its predecessor from 2023. Similar to METRICX-23, it is based on the mT5-XXL pretrained language model, which is fine-tuned in two stages on DA and MQM scores from WMT 2015-22, and it implements three major design improvements. First, the training data in both stages is augmented with synthetic examples to make the metric more robust to several common failure modes, such as fluent but unrelated translation, or undertranslation. Second, a small proportion of DA data is mixed in during the second stage of fine-tuning in order to preserve the performance on non-MQM language pairs. Finally, the model's training is done on a mixture of examples that include the source only, the reference only, or both, which allows the model to operate in both a QE and a reference-based mode (and the latter either with or without the source included). Hence, both METRICX-24-HYBRID and METRICX-24-HYBRID-QE submission are in fact the exact same model, only with the references excluded from the input in the latter case.

SENTINEL-CAND-MQM, SENTINEL-REF-MQM and SENTINEL-SRC-MQM (Perrella et al., 2024) are designed explicitly to scrutinize the accuracy, robustness, and fairness of the meta-evaluation process. The three sentinel metrics are trained only on the candidate, reference and source sentence respectively on DA and MQM data from WMT 2017 to 2022.

XCOMET AND XCOMET-QE (Guerreiro et al., 2023) models are trained using both a sentence-level signal and span-level supervision coming from MQM data from previous years, along with some synthetic data that mimics hallucinations. We ensemble XCOMET-XXL and XCOMET-XL to give a single unified score.

XLSIMMQM (Mukherjee and Shrivastava, 2023b) is an enhanced version of XLSIM, a supervised reference-based evaluation metric, which we have transformed into a reference-free model to improve its applicability across multiple language pairs. Unlike the original XLSIM, which was limited to the English-German language pair, XLSIMMQM is trained on a filtered comprehensive dataset curated from WMT-MQM (2020-22), ensuring broader applicability and robustness. The filtered datasets (train, dev and test) contains uniform distribution across good, medium and poor-quality sentences; this careful balancing of the dataset leads to a better, reliable and robust metric.

5 Meta Evaluation

The goal of metric meta-evaluation is to quantify how well automatic metrics agree with human ratings of translation quality. There are a multitude of ways to approach this problem, as evidenced by the variety of solutions proposed by previous years' editions of the shared task. For instance—to name just a few possible design decisions—the agreement can be measured at the system or segment level; the agreement function can be Pearson, Spearman, Kendall, pairwise agreement, or L_2 loss; the agreement can be computed per domain or on the full dataset. None of these approaches are necessarily right or wrong, but rather each method evaluates a different property of the metric.

Because there is no one way to evaluate a metric, the past two iterations of the Metrics Shared Task defined a variety of "tasks" (or different configurations of meta-evaluations) that evaluated some aspect of a metric, then calculated an overall quality score by averaging the individual task scores. Implicitly, this approach defines a "high-quality" metric as one that performs well across the tasks on average. In 2022, there were 201 tasks that varied along dimensions such as language pair, domain, correlation granularity, correlation statistic, etc. In

2023, the number of tasks was reduced to 10, measuring only pairwise accuracy and Pearson at both the system and segment levels.

For this year's meta-evaluation, we follow the same approach of averaging performance across tasks, but focus the tasks to better align with how evaluation metrics are used in practice. The two main use cases that we targeted were using metrics to rank a set of MT systems and using a metric to rank a set of translations for the same source segment. The former setting is widely used by academics and practitioners in industry to determine whether one model produces better translations than another, and the latter setting has applications in Minimum Bayes Risk Decoding and Quality Estimation Reranking either directly as decoding method (Fernandes et al., 2022; Freitag et al., 2022) or to further fine-tune models (Finkelstein and Freitag, 2024; Finkelstein et al., 2024). The latter one is getting more popular and can introduce metric biases (Kovacs et al., 2024) that is an emerging challenge for metrics. As such, we defined one task to quantify how well metrics work for each of these two use cases separately for all three language pairs, resulting in a total of six tasks.

At the system-level, we use the recently proposed metric called soft pairwise accuracy, or SPA (Thompson et al., 2024). One of the drawbacks of standard pairwise accuracy (or the very related Kendall's τ) that has been used in previous years' shared tasks is that it does not account for the uncertainty of the system ranking. For example, if the human ranking of two systems is almost arbitrary (e.g, a statistical tie) but the metric ranking is quite certain, standard pairwise accuracy will either reward or penalize the metric nearly randomly. The reverse case—a certain human ranking and uncertain metric ranking—also nearly arbitrarily rewards or penalizes metrics. If both rankings are uncertain, the metric will again be rewarded nearly randomly, and the penalty for an incorrect ranking is equal to when the metric was very certain but also wrong.

SPA addresses this problem by using p-values as a proxy for certainty, calculating p-values between two systems using both the metric and human scores, then taking 1.0 minus the absolute difference between the two p-values as the metric's score for that pair. This rewards metrics that result in the same statistical conclusion as the human scores. Now, statistical ties do not randomly reward or penalize metrics, but instead the score is proportional to whether or not the metric and human have the same level of certainty in the ranking.

language	ref used	scored ref
en→de	B	A
en→es	A	–
ja→zh	A	–

Table 7: Use of reference translations.

task	lang	level	correlation	wt
1	en→de	system	SPA	1
2	en→de	segment	acc^*_{eq}	1
3	en→es	system	SPA	1
4	en→es	segment	acc^*_{eq}	1
5	ja→zh	system	SPA	1
6	ja→zh	segment	acc^*_{eq}	1

Table 8: For each language pair, soft pairwise accuracy (SPA) was used at the system-level and acc^*_{eq} at the segment-level. Each task was given equal weight in the overall average. See §5 for explanations of SPA and acc^*_{eq}.

At the segment-level, we follow last year's meta-evaluation and meta-evaluate metrics using "group-by-item" segment-level accuracy with tie calibration (Deutsch et al., 2023) denoted acc^*_{eq}.

The six tasks (shown in Table 8) receive equal weighting in the overall average, which is the final score for the metric.

Removing Pearson's Correlation: Notably, the meta-evaluation this year only focuses on evaluating rankings and does not include any correlation that evaluates the absolute value of the scores predicted by metrics, like Pearson's correlation. This decision was made because using metrics to rank systems or translations is much more common in practice than using a metric to approximate the absolute quality score as derived by humans, which is more similar to a Pearson correlation.

Limitations: Like previous years, 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, which could result in certain tasks carrying more weight than others in the overall ranking. However, to simplify the implementation, we assigned equal weight to all tasks, which worked well in last year's evaluation.

5.1 Rank Assignment

For each task, we assign ranks to metrics based on their significance clusters in the same way that we did last year, detailed below.

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 resampling 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 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 ith 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.

5.2 Implementation Details

The code for running the meta-evaluation is available in the MT Metrics Eval library.[11]

To calculate p-values for SPA, we use a paired permutation test (Noreen, 1989) with 1k resamples.

In previous years' shared tasks, tasks were categorized based on whether they included additional reference translations in the overall system ranking. Following last year's proposal, we always include the additional reference in the overall ranking. This year, this only applies to en→de which is the only language pair with more than one reference translation (see Table 7).

Out of all the submitted MT systems, MSLC consistently scores well below the other systems for all language pairs and was identified as an outlier and removed from the correlation calculation.

6 Main Results

As we have described in Section 5, the final statistic used to rank the metrics is defined as the average of the results from the six main tasks (system-level and segment-level tasks in different language pairs). Table 1 shows the official scores and rankings of all baselines and primary submissions. Table 9 shows 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 26 evaluated metrics, BLEU, SPBLEU and CHRF are ranked 23rd, 22nd and 20th respectively. Fine-tuned neural metrics, like XCOMET and METRICX-23 are the highest ranked non-ensemble metrics. The ensemble submission METAMETRIC_MT is in the same significance cluster as XCOMET and METRICX-24-HYBRID, but relies heavily on the 2023 version of METRICX-24-HYBRID. Like last year, QE metrics perform very well, with METRICX-24-HYBRID-QE and GEMBA_ESA sharing the second significance cluster.

Figure 1 shows the correlation scores split by language pair. Interestingly, GEMBA_ESA is performing very well for en→es and ja→zh, while ranked below many metrics for en→de. GEMBA_ESA is

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

Metric	avg-corr		en-de sys SPA task1		en-de seg acc^{*}_{eq} task2		en-es sys SPA task3		en-es seg acc^{*}_{eq} task4		ja-zh sys SPA task5		ja-zh seg acc^{*}_{eq} task6	
MetaMetrics-MT	**1**	**0.725**	2	0.883	**1**	**0.542**	**1**	**0.804**	2	0.686	2	0.873	**1**	**0.561**
MetricX-24-Hybrid	**1**	**0.721**	2	0.874	2	0.532	2	0.799	3	0.685	**1**	**0.897**	2	0.539
XCOMET	**1**	**0.719**	**1**	**0.905**	2	0.530	2	0.791	**1**	**0.688**	**1**	**0.890**	5	0.510
MetricX-24-Hybrid-QE*	2	0.714	2	0.878	3	0.526	2	0.789	4	0.685	2	0.875	3	0.530
gemba_esa*	2	0.711	4	0.793	5	0.507	**1**	**0.838**	5	0.683	**1**	**0.908**	2	0.539
XCOMET-QE*	3	0.695	**1**	**0.889**	4	0.520	**1**	**0.801**	2	0.687	4	0.808	10	0.463
COMET-22	3	0.688	2	0.879	8	0.482	2	0.778	5	0.683	4	0.813	6	0.496
BLEURT-20	3	0.686	2	0.881	7	0.486	3	0.695	6	0.681	**1**	**0.887**	8	0.484
MetaMetrics-MT-QE*	3	0.684	2	0.860	6	0.497	3	0.711	2	0.686	3	0.837	4	0.516
bright-qe*	4	0.681	3	0.816	6	0.500	2	0.792	**1**	**0.689**	4	0.805	8	0.484
BLCOM_1	4	0.664	3	0.840	10	0.455	3	0.680	6	0.681	3	0.843	7	0.488
sentinel-cand-mqm*	5	0.650	3	0.822	4	0.517	2	0.785	4	0.683	7	0.610	8	0.481
PrismRefMedium	5	0.646	4	0.776	14	0.434	3	0.652	7	0.680	2	0.872	10	0.462
PrismRefSmall	5	0.642	4	0.772	14	0.433	4	0.634	8	0.680	2	0.875	11	0.457
CometKiwi*	5	0.640	5	0.732	9	0.467	3	0.693	4	0.684	5	0.776	7	0.490
damonmonli	5	0.635	5	0.696	12	0.443	4	0.607	6	0.682	**1**	**0.911**	9	0.472
YiSi-1	6	0.630	4	0.759	13	0.436	4	0.609	7	0.681	3	0.835	11	0.458
BERTScore	7	0.617	4	0.749	14	0.435	4	0.587	6	0.682	4	0.799	12	0.451
MEE4	7	0.609	5	0.731	13	0.437	7	0.504	4	0.683	2	0.855	13	0.446
chrF	8	0.608	4	0.750	15	0.431	5	0.581	8	0.680	5	0.767	16	0.436
chrfS	8	0.606	4	0.742	14	0.434	6	0.549	6	0.682	4	0.788	14	0.444
spBLEU	9	0.593	4	0.741	17	0.431	6	0.523	7	0.680	6	0.744	16	0.436
BLEU	9	0.589	4	0.736	16	0.431	6	0.512	8	0.680	6	0.740	17	0.435
XLsimMqm*	10	0.515	6	0.612	11	0.450	8	0.359	7	0.681	7	0.548	15	0.438
sentinel-src-mqm*	10	0.513	7	0.406	18	0.429	5	0.580	8	0.680	8	0.546	17	0.435
sentinel-ref-mqm	10	0.513	7	0.405	18	0.429	4	0.581	8	0.680	8	0.545	17	0.435

Table 9: Correlation results per task for the main language pairs. See §5 for descriptions of soft pairwise accuracy (SPA) and acc^{*}_{eq}. Rows are sorted by the overall average correlation across all 6 tasks (leftmost column). Starred metrics are reference-free, and underlined metrics are baselines.

a prompt-based metric and not fine-tuned for any metric task. Both en→es and ja→zh are new language pairs, and no fine-tuning data exists which might have played in disadvantage for all fine-tuned metrics.

We continue to be interested in metrics' abilities to generalise across domains. In Figure 2, we present the performance of each metric across different domains. Similar to last year, we observe that neural metrics perform better than lexical overlap metrics across all four domains. Figure 3 shows the average correlations of metrics when grouped separately by system-level and segment-level tasks. There is a high correlation between the rankings of both granularities.

7 Beyond accuracy and correlation

Last year, we conducted two additional analyses beyond correlation with human scores to find the threshold of metrics' score differences correspond to statistical significance of MT system rankings demonstrated by human annotators and the metrics themselves. Despite the better correlation with human judgements achieved by new neural metrics, BLEU remains as the most used metric in the MT research community. One of the reasons is that MT researchers have established some "shared understanding" about the relationship between BLEU and the actual translation quality, and similar intuitions about new metrics have yet to crystallize. Our analyses beyond correlation provided an interpretation of the metrics' score differences. Hence, we are continuing such analyses to support building an intuitive sense of metric score meanings and encourage broader adoption of new automatic MT evaluation metrics. As a reminder, our results should *NOT* be used as arguments to forego significance tests or appropriate human evaluation.

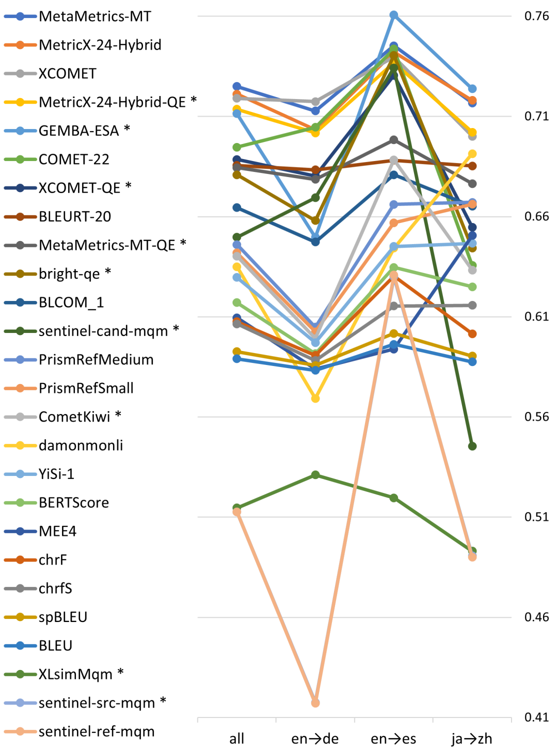

Figure 1: Average metrics' meta-evaluation scores in tasks grouped by language pair.

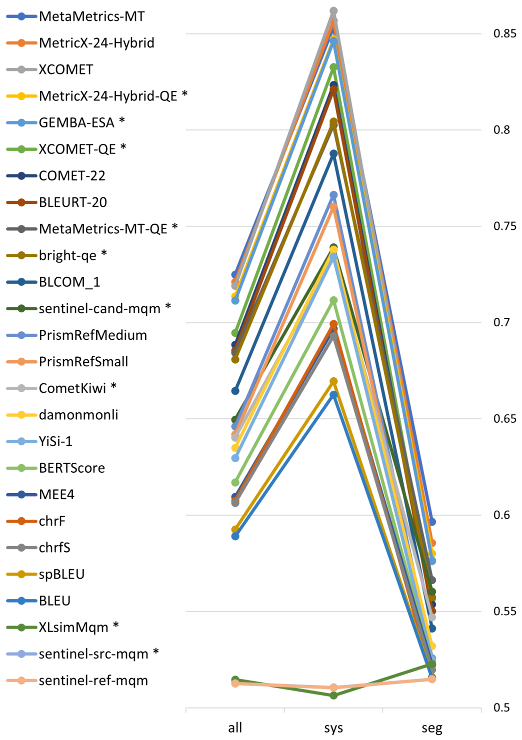

Figure 3: Average metrics' correlation with human in tasks grouped by granularity level.

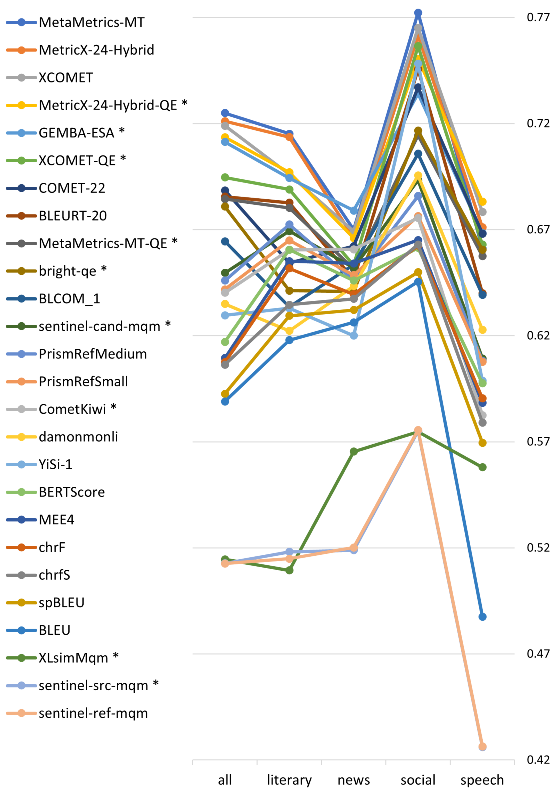

Figure 2: Average metrics' correlation with human in tasks grouped by domain.

7.1 Correspondence to MQM scores significance

We first study the relationship between statistically significant differences in human scores and the magnitude of metric differences as in Lo et al. (2023a). We run a two-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.[12] We set $p_{mqm} < 0.05$ as the significance level of MQM 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 ΔM is the metric score difference.

Figure 4 shows the (log) p-value of two-sided paired t-test on the MQM scores against the corre-

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

sponding BLEU and COMET-22 score difference for each system pair in en→de. Figures 6-10 in appendix D, show the same analyses for all metrics and language pairs. 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. 5.4 for BLEU in Figure 4. Using this lookup method, Table 10 shows the cut-offs of ΔM when $Pr(p_{mqm} < 0.05|\Delta M) = 0.8$ for each metric and language pair.

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, except for BLEU, BRIGHT-QE, COMETKIWI, MEE4, METAMETRICS_MT_MQM_QE_KENDALL.SEG.S, SPBLEU and XLSIMMQM. This means the metric cut-offs we find using the regression model are reliable.

Contrary to the shared understanding that 2 BLEU improvement represents "significant" or "notable by human" improvement in the actual translation quality, our analyses show that 5.4 BLEU improvement is required to be confident (80%) that the MQM scores would be different with statistical significance for en→de and that threshold would be as high as 11 BLEU for en→es. 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 ja→zh translation quality, we see that a BLEU difference of 1.4 corresponds to 80% confidence that the metric's ranking of the two MT systems will match the decision made by human annotators with a significant difference. Meanwhile, a COMET-22 score difference of 0.021 would have the same 80% chance of human judged significant difference.

7.2 Correspondence to metric scores significance

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 as inspired by Marie (2022). Instead of the two-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. 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 ΔM is the metric score difference.

Figure 5 shows the (log) p-value of the significance test with bootstrap resampling on the metric scores for BLEU and COMET-22 score difference of each system pair in en→de. Additional figures (Figures 11-15 in appendix Appendix D) show the same analyses for all metrics and language pairs. Using the same lookup method described in the previous subsection, Table 11 shows the cut-offs of ΔM when $Pr(p_M < 0.05|\Delta M) = 0.8$ for each metric and language pair.

We run the leave-one-system-out cross validation, 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.

Table 11 serves as a reference of metric differences that correspond to statistical significance with high confidence. For example, when evaluating en→de translation quality, we see that a BLEU difference of 0.97 corresponds to 80% confidence the difference is statistically significant. Meanwhile, a COMET-22 score difference of 0.0043 would have the same 80% chance of statistical significance. Our results, agreeing with Marie (2022), show that to claim significant differences ($p_M < 0.05$) in BLEU with high confidence (80%), the differences should be much higher than the shared understanding of 0.5 BLEU, ranging from 0.89 to 0.97 for the three language pairs.

Closely related to this analysis, Kocmi et al. (2024b) investigated the agreement between human evaluations and metric differences, employing pairwise accuracy as the meta-evaluation metric. Assuming an 80% agreement rate with human judgments, their findings align closely with ours for pretrained metrics but not for metrics such as BLEU or ChrF. For instance, COMET-22 requires a score difference of 0.0056 to achieve 80% accuracy with humans, compared to our range of 0.0043–0.0055. Similarly, CometKiwi requires a

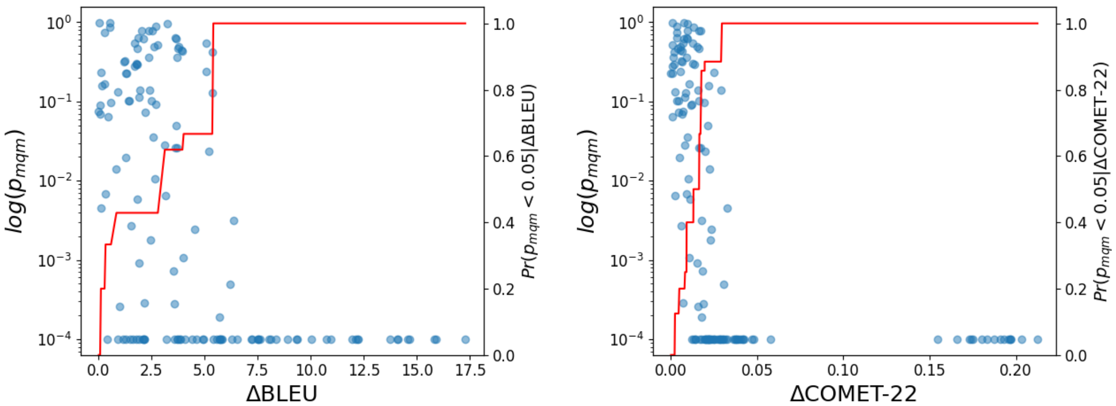

Figure 4: Log p-value of two-sided paired t-test on MQM scores (p_{mqm}) against the metric (left: BLEU, right: COMET-22) score difference for each system pair in en→de. The red line is the isotonic regression fit to all data points, representing $Pr(p_{mqm} < 0.05|\Delta M)$. Note: for readability, values of p_{mqm} are rounded up to 0.0001 when they are less than 0.0001.

	en→de		en→es		ja→zh	
Metric	min ΔM	c.v. precision	min ΔM	c.v. precision	min ΔM	c.v. precision
BERTSCORE	0.0099	[50-100%]	0.018	[50-100%]	0.013	[64-100%]
BLCOM_1	0.022	[75-100%]	0.034	[50-100%]	0.021	[62-100%]
BLEU	5.4	[67-100%]	11	[0-100%]	1.4	[50-100%]
BLEURT-20	0.021	[62-100%]	0.014	[60-100%]	0.029	[80-100%]
BRIGHT-QE	0.018	[20-100%]	0.049	[50-100%]	0.061	[62-100%]
CHRF	3.0	[67-100%]	2.1	[57-100%]	3.5	[78-100%]
CHRFS	0.023	[50-100%]	0.043	[50-100%]	0.021	[60-100%]
COMET-22	0.018	[50-100%]	0.017	[60-100%]	0.021	[60-100%]
COMETKIWI	0.024	[17-100%]	0.027	[33-100%]	0.050	[67-100%]
DAMONMONLI	0.84	[27-100%]	0.064	[50-100%]	0.51	[88-100%]
GEMBA_ESA	4.5	[70-100%]	1.5	[67-100%]	4.8	[86-100%]
MEE4	0.019	[25-100%]	0.028	[33-100%]	0.019	[55-100%]
metametrics_mt_mqm_hybrid_kendall	0.029	[53-100%]	0.066	[60-100%]	0.066	[70-100%]
metametrics_mt_mqm_qe_kendall.seg.s	0.016	[14-100%]	0.025	[50-100%]	0.031	[67-100%]
METRICX-24-HYBRID	0.52	[73-100%]	0.95	[62-100%]	0.60	[75-100%]
METRICX-24-HYBRID-QE	0.44	[62-100%]	0.39	[67-100%]	0.63	[78-100%]
PRISMREFMEDIUM	0.073	[67-100%]	0.12	[50-100%]	0.14	[56-100%]
PRISMREFSMALL	0.10	[67-100%]	0.15	[50-100%]	0.15	[56-100%]
SENTINEL-CAND-MQM	0.066	[50-100%]	0.13	[50-100%]	0.088	[55-100%]
SENTINEL-REF-MQM	—	—	—	—	—	—
SENTINEL-SRC-MQM	—	—	—	—	—	—
SPBLEU	4.3	[50-100%]	9.1	[0-100%]	4.0	[75-100%]
XCOMET	0.022	[53-100%]	0.025	[67-100%]	0.046	[78-100%]
XCOMET-QE	0.013	[50-100%]	0.029	[50-100%]	0.062	[67-100%]
XLSIMMQM	0.018	[100-100%]	0.0012	[57-100%]	0.004	[43-100%]
YISI-1	0.0063	[60-100%]	0.0098	[56-100%]	0.012	[75-100%]

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

difference of 0.0053, while our results range from 0.0037 to 0.0056. Conversely, for BLEU, their analysis suggests an expected improvement of 2.34 BLEU points for 80% agreement, whereas our analysis indicates a need for an improvement of 0.89–0.97 BLEU points. However, it is important to note that we are comparing distinct metrics, and that confidence levels are not directly comparable to agreement rates.

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 lexical metrics towards more recent and stronger metrics.

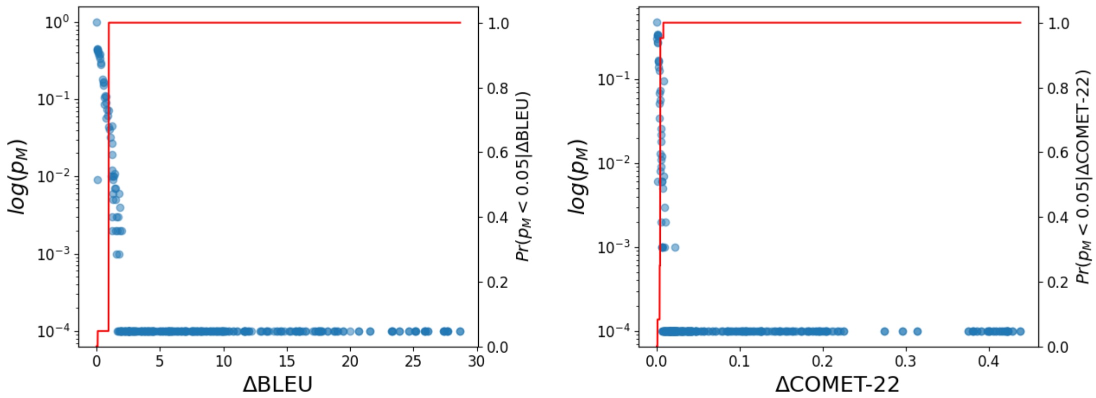

Figure 5: Log p-value of significance test with bootstrap resampling (p_M) on system-level metric scores against each metric (left: BLEU, right: COMET-22) score difference for each system pair in en→de. The red line is the isotonic regression fit to all data points, representing $Pr(p_M < 0.05|\Delta M)$. Note: for readability, values of p_M are rounded up to 0.0001 when they are less than 0.0001.

Metric	en→de min ΔM	en→de c.v. precision	en→es min ΔM	en→es c.v. precision	ja→zh min ΔM	ja→zh c.v. precision
BERTSCORE	0.0028	[92-100%]	0.0028	[100-100%]	0.0044	[100-100%]
BLCOM_1	0.0039	[100-100%]	0.0055	[100-100%]	0.0044	[100-100%]
BLEU	0.97	[100-100%]	0.93	[100-100%]	0.89	[91-100%]
BLEURT-20	0.0056	[96-100%]	0.0053	[94-100%]	0.0068	[95-100%]
BRIGHT-QE	0.0041	[89-100%]	0.0078	[94-100%]	0.024	[95-100%]
CHRF	0.83	[96-100%]	0.77	[94-100%]	0.89	[100-100%]
CHRFS	0.0051	[91-100%]	0.0054	[95-100%]	0.0055	[95-100%]
COMET-22	0.0043	[96-100%]	0.0055	[86-100%]	0.0046	[95-100%]
COMETKIWI	0.0037	[100-100%]	0.0048	[82-100%]	0.0056	[100-100%]
DAMONMONLI	0.20	[94-100%]	0.17	[82-100%]	0.41	[90-100%]
GEMBA_ESA	0.82	[92-100%]	0.85	[91-100%]	1.4	[100-100%]
MEE4	0.0042	[95-100%]	0.0051	[86-100%]	0.0057	[95-100%]
metametrics_mt_mqm_hybrid_kendall	0.0067	[92-100%]	0.0081	[89-100%]	0.013	[90-100%]
metametrics_mt_mqm_qe_kendall.seg.s	0.0038	[89-100%]	0.0050	[80-100%]	0.0089	[95-100%]
METRICX-24-HYBRID	0.11	[100-100%]	0.15	[100-100%]	0.14	[95-100%]
METRICX-24-HYBRID-QE	0.087	[90-100%]	0.14	[100-100%]	0.12	[100-100%]
SENTINEL-CAND-MQM	0.011	[96-100%]	0.013	[95-100%]	0.030	[95-100%]
SENTINEL-REF-MQM	—	—	—	—	—	—
SENTINEL-SRC-MQM	—	—	—	—	—	—
SPBLEU	0.96	[96-100%]	1.1	[95-100%]	1.0	[100-100%]
PRISMREFMEDIUM	0.019	[95-100%]	0.02	[100-100%]	0.036	[90-100%]
PRISMREFSMALL	0.023	[96-100%]	0.022	[100-100%]	0.042	[95-100%]
XCOMET	0.0051	[100-100%]	0.0065	[86-100%]	0.010	[95-100%]
XCOMET-QE	0.0044	[96-100%]	0.0058	[94-100%]	0.0099	[100-100%]
XLSIMMQM	0.0036	[82-100%]	0.0013	[90-100%]	0.0019	[79-100%]
YISI-1	0.0010	[91-100%]	0.0014	[90-100%]	0.0051	[100-100%]

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

8 ESA 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 WMT24 General MT Shared Task (Kocmi et al., 2024a), which ran human evaluation for nine language pairs.

In contrast to previous years, WMT24 redefined their human evaluation process and developed a new method called Error Span Analysis (ESA, Kocmi et al. (2024c)), a method that simplifies MQM by asking annotators only to mark error spans and classify them either as minor or major severity. In addition to that, the annotator is asked to mark the whole segment with a score of 0–100 in the SQM fashion. As Kocmi et al. (2024c) claim, the method is cheaper than MQM to annotate, yet it produces closer human judgment to MQM annotations than the formerly used DA+SQM (Kocmi et al., 2023) due to being less affected by fluency.

We present system-level accuracy results for both MQM and ESA in Table 15. There are many factors that could affect the ranking. Apart from using a different human annotation protocol, MQM compares 3 language pairs whereas ESA compares 9 language pairs, containing also two low-resource pairs: Czech→Ukrainian and English→Icelandic. There is an overlap of only one language pair between the two: English→Spanish.

Most of the metrics have a similar ranking for both MQM and ESA; however, there are two metrics with largely different rankings: GEMBA_ESA and metametrics_mt_mqm_qe_kendall.seg.s, whose rankings are significantly lower under ESA than for MQM. The likely explanation for GEMBA_ESA is that ESA doesn't produce ties, in contrast to MQM, whereas GEMBA_ESA produces them regularly. As for the latter metric, we don't see any clear pattern except for having low performance for Czech→Ukrainian.

9 Challenge Sets Sub-task

For the third year, the Metrics Shared Task included a sub-task involving 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 is conducted 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 or evaluating particular translation phenomena. In order to shed light on different perspectives on evaluation, the sub-task takes place in a decentralized manner, since 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 examples for different phenomena, where every example $(s, t, r) \in S$ contains one source sentence s, one translation hypothesis t and one reference r.

2. In the *Scoring Round*, The metrics participants from the main task (the *Builders*) are asked to score with their metrics the translations in the given test set. Also, in this phase, the metrics task 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 employ their own evaluation for the performance of the metrics with regard to the phenomena they intended to test.

This year there were 4 submissions, covering a wide range of phenomena and 23 different language pairs, which supersede the official language pairs of the Metrics Shared Task. An overview of the submitted challenge sets can be seen in Table 12. A short description of every submission follows:

AfriMTE Challenge Set The AfriMTE challenge set (Wang et al., 2024b) aims to evaluate the capabilities of metrics for machine translation on low-resource languages, primarily assessing cross-lingual transfer learning and generalization across a wide range of under-resourced African languages. The challenge set concentrates on the subsets of the FLORES-200 dataset (NLLB-Team et al., 2022) and covers 13 language pairs. Specifically, there are Darija-French, English-Egyptian Arabic, English-French, English-Hausa, English-Igbo, English-Kikuyu, English-Luo, English-Somali, English-Swahili, English-Twi, English-isiXhosa, English-Yoruba, and Yoruba-English. Originally, AfriMTE (Wang et al., 2024a) provides both fine-grained word-level error annotations and sentence-level Direct Assessment scoring for translation adequacy and fluency. For this year's challenge set sub-task, we utilize the translation adequacy test set from AfriMTE as the African Challenge set to evaluate the sentence-level scoring performance of metrics. The analysis of the task submissions (Wang et al., 2024b) has yielded several insights. First, language-specific adaptation, cross-lingual transfer learning, and larger language model sizes significantly enhance metric performance. Second, moderately-sized supervised models can attain robust performance when augmented with language adaptation techniques tailored to

Challenge Set	Directions	Phenomena	Items	Citation	Link (https://github.com/)
AfriMTE	13	African languages	2,815	Wang et al. (2024b)	masakhane-io/africomet
BioMQM	11	biomedical domain	4,641	Zouhar et al. (2024)	thompsonb/bio-mqm-dataset
DFKI	2	linguistic phenomena	137,000	Avramidis et al. (2024)	DFKI-NLP/mt-testsuite
MSLC24	3	low quality MT	964	Knowles et al. (2024)	nrc-cnrc/MSLC

Table 12: Overview of the participation at the metrics challenge sets sub-task.

low-resource African languages during pretraining. Last, submissions demonstrate promising outcomes for language pairs such as Darija-French, English-Egyptian Arabic, and English-Swahili. However, considerable challenges remain for extremely low-resource languages like English-Luo and English-Twi, underscoring critical areas for future research and improvement in machine translation metrics for African languages.

BioMQM Recent work (Zouhar et al., 2024) has compared trained versus untrained metric performance on the WMT domains compared to the biomedical domain and shown that trained metrics appear to be over-fitting on the domains used in the WMT Metrics Shared Tasks. This is likely due to trained metrics using prior WMT metrics datasets, and then being evaluated on very similar data in the latest WMT Metrics Shared Task. Zouhar et al. (2024) released a biomedical dataset (BioMQM) consisting of source sentences and translations from Yeganova et al. (2021) along with new translations and MQM annotations. We produce scores on the BioMQM for the latest metrics (all those submitted to this Metrics Shared Task, plus the baseline metrics) and release them for future analysis.[13]

DFKI Challenge Set This year's submission by DFKI (Avramidis et al., 2024) expands the linguistically motivated challenge set of previous years (Avramidis et al., 2023; Avramidis and Macketanz, 2022), including 137,000 items in overall, extracted from 100 MT systems for the two language directions (en→de, en→ru), covering more than 100 linguistically-motivated phenomena organized in 14 linguistic categories. The metrics with the statistically significant best performance with regard to our linguistically motivated analysis are METRICX-24-HYBRID and METRICX-24 for en→de and METRICX-24 for en→ru, whereas METAMETRICS and XCOMET are in the next ranking positions in both language pairs. Metrics are more accurate in detecting linguistic errors among LLM translations than in translations based on the encoder-decoder NMT architecture. Some of the most difficult phenomena for the metrics to score are the transitive past progressive, the multiple connectors, the ditransitive simple future I for en→de and pseudogapping, contact clause and cleft sentences for en→ru. The LLM-based metric GEMBA, despite the overall low performance, has the best performance on scoring German negation errors.

MSLC24 Challenge Set Building on the Metric Score Landscape Challenge (MSLC23; Lo et al., 2023b), which aims to provide a view of metric performance on a broader range of MT quality, MSLC24 includes a collection of low- to medium-quality MT systems' output on the news portion of the WMT24 General MT Shared Task test set, as well as some specific phenomena that may result in unexpected behaviors from some metrics, such as empty strings in source/reference/hypothesis, wrong/mixed language output and different language variants. MSLC24 focuses on three language pairs (English→German, English→Spanish and Japanese→Chinese). The authors also submit the top system in this challenge set to the General Translation task in order to obtain human evaluation. Together with the high quality systems by other participants submitted to the General MT Shared Task, this enables better interpretation of metric scores across a range of different levels of translation quality and analyse metric characteristics beyond just correlation. The results of MSLC24 highlight the importance of examining real-word corner cases and issues of reproducibility in order to more responsibly introduce new metrics to the research community.

10 Conclusion

This paper summarizes the results of the WMT24 shared task on automated machine translation evaluation, the Metrics Shared Task. We presented an extensive analysis on how well metrics perform on

[13] https://github.com/thompsonb/bio-mqm-dataset/tree/main/data/WMT24_Metrics_ChallengeSet

our three main language pairs: English→German, English→Spanish and Japanese→Chinese. The results, based on 6 different tasks, confirm the superiority of neural-based learned metrics over overlap-based metrics like BLEU, SPBLEU or CHRF. These results are confirmed with ESA human judgement. Overall, we did not find any issues for neural fine-tuned metrics when evaluating LLM-based translations. 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.

11 Ethical Considerations

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

Organizers from the National Research Council Canada, 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 MetricX 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.

12 Acknowledgments

Results for this shared task would not be possible without tight collaboration with the organizers of the WMT24 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.

Brian Thompson's work on this shared task is unrelated to and conducted independently of the author's position at Amazon. David Adelani is supported by Canada CIFAR AI Chair program.

Chrysoula Zerva's work is supported by the Portuguese Recovery and Resilience Plan through project C645008882- 00000055 (NextGenAI - Center for Responsible AI), by the EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

References

David Anugraha, Garry Kuwanto, Lucky Susanto, Derry Tanti Wijaya, and Genta Indra Winata. 2024. Metametrics-MT: Tuning machine translation metametrics via human preference calibration. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Eleftherios Avramidis and Vivien Macketanz. 2022. Linguistically motivated evaluation of machine translation metrics based on a challenge set. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 514–529, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eleftherios Avramidis, Shushen Manakhimova, Vivien Macketanz, and Sebastian Moeller. 2024. Machine translation metrics are better in evaluating linguistic errors on llms than on encoder-decoder systems. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, Florida, USA. Association for Computational Linguistics.

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*, pages 713–729, Singapore. Association for Computational Linguistics.

Yanran Chen and Steffen Eger. 2023. MENLI: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

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: Meta-Evaluating Modern Metrics with Pairwise Accuracy and Tie Calibration. pages 12914–12929.

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.

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

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

Mara Finkelstein, David Vilar, and Markus Freitag. 2024. Introducing the newspalm mbr and qe dataset: Llm-generated high-quality parallel data outperforms traditional web-crawled data. In *Proceedings of the Ninth Conference on Machine Translation*, USA. 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, 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.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

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*.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Rebecca Knowles, Samuel Larkin, and Chi-kiu Lo. 2024. MSLC24: Further challenges for metrics on a wide landscape of translation quality. In *Proceedings of the Ninth Conference on Machine Translation*, USA. 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, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinþór Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task:
the LLM era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, USA. 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 and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

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

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024c. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

Geza Kovacs, Daniel Deutsch, and Markus Freitag. 2024. Mitigating metric bias in minimum bayes risk decoding. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

AI @ Meta Llama Team. 2024. The llama 3 herd of models.

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, Macau SAR, China. Asia-Pacific Association for Machine Translation.

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*, pages 776–799, 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.

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. 2023a. MEE4 and XLsim : IIIT HYD's submissions' for WMT23 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 800–805, Singapore. Association for Computational Linguistics.

Ananya Mukherjee and Manish Shrivastava. 2023b. 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.

Ananya Mukherjee and Manish Shrivastava. 2024. chrf-s: Semantics is all you need. In *Proceedings of the Ninth Conference on Machine Translation*, USA. Association for Computational Linguistics.

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

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*.

Eric W Noreen. 1989. Computer intensive methods for hypothesis testing: An introduction. *Wiley, New York*, 19:21.

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

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

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.

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

Parker Riley, Daniel Deutsch, George Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus Freitag. 2024. Finding replicable human evaluations via stable ranking probability. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4908–4919, Mexico City, Mexico. Association for Computational Linguistics.

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

Davor Runje and Sharath M Shankaranarayana. 2023. Constrained monotonic neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29338–29353. PMLR.

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, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy.

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.

Jiayi Wang, David Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Mohamed, Hassan Ayinde, Oluwabusayo Awoyomi, Lama Alkhaled, Sana Al-azzawi, Naome Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Toadoum Sari Sakayo, Lyse Naomi Wamba, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Iro, Saheed Abdullahi, Stephen Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Ogbu, Sam Ochieng', Verrah Otiende, Chinedu Mbonu, Yao Lu, and Pontus Stenetorp. 2024a. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

Jiayi Wang, David Ifeoluwa Adelani, and Pontus Stenetorp. 2024b. Evaluating WMT 2024 metrics shared task submissions on afrimte (the african challenge set). In *Proceedings of the Ninth Conference on Machine Translation*, Miami, Florida, USA. 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.

Genta Indra Winata, David Anugraha, Lucky Susanto, Garry Kuwanto, and Derry Tanti Wijaya. 2024. Metametrics: Calibrating metrics for generation tasks using human preferences. *arXiv preprint arXiv:2410.02381*.

Lana Yeganova, Dina Wiemann, Mariana Neves, Federica Vezzani, Amy Siu, Inigo Jauregi Unanue, Maite Oronoz, Nancy Mah, Aurélie Névéol, David Martinez, Rachel Bawden, Giorgio Maria Di Nunzio, Roland Roller, Philippe Thomas, Cristian Grozea, Olatz Perez-de Viñaspre, Maika Vicente Navarro, and Antonio Jimeno Yepes. 2021. Findings of the WMT 2021 biomedical translation shared task: Summaries of animal experiments as new test set. In *Proceedings of the Sixth Conference on Machine Translation*, pages 664–683, Online. Association for Computational Linguistics.

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*.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500, Bangkok, Thailand. Association for Computational Linguistics.

A Correlations with MQM for all metrics

Table 13 contains the results for all metrics (including contrastive submissions) on the 6 standard tasks described in Table 8.

Metric	avg-corr		en-de sys SPA task1		en-de seg acc^*_{eq} task2		en-es sys SPA task3		en-es seg acc^*_{eq} task4		ja-zh sys SPA task5		ja-zh seg acc^*_{eq} task6	
MetricX-24	**1**	**0.725**	2	0.873	2	0.534	2	0.789	3	0.685	**1**	**0.921**	2	0.547
MetaMetrics-MT	**1**	**0.725**	2	0.882	**1**	**0.542**	2	0.805	2	0.686	3	0.872	**1**	**0.561**
metametrics_mt_mqm_kendall	**1**	**0.724**	2	0.882	**1**	**0.542**	2	0.804	2	0.686	3	0.871	**1**	**0.561**
metametrics_mt_mqm_same_source_targ	2	0.723	**1**	**0.883**	**1**	**0.542**	2	0.803	2	0.686	3	0.874	2	0.550
MetricX-24-Hybrid	2	0.720	2	0.873	2	0.532	2	0.796	3	0.685	2	0.895	3	0.539
XCOMET	2	0.719	**1**	**0.906**	3	0.530	2	0.788	**1**	**0.688**	2	0.890	7	0.510
MetricX-24-Hybrid-QE*	3	0.714	2	0.880	4	0.526	2	0.790	4	0.685	3	0.875	4	0.530
gemba_esa*	3	0.712	4	0.793	6	0.507	**1**	**0.838**	5	0.683	**1**	**0.909**	3	0.539
*MetricX-24-QE**	3	0.710	2	0.880	3	0.528	3	0.772	3	0.685	3	0.875	5	0.522
*CometKiwi-XXL**	3	0.703	3	0.839	9	0.481	**1**	**0.843**	8	0.680	2	0.881	8	0.494
XCOMET-QE*	4	0.695	**1**	**0.890**	5	0.520	2	0.801	2	0.687	5	0.809	12	0.463
COMET-22	4	0.689	2	0.877	9	0.482	2	0.782	5	0.683	5	0.815	8	0.496
*metametrics_mt_mqm_qe_same_source_t**	4	0.688	2	0.860	7	0.497	4	0.709	2	0.686	4	0.853	5	0.524
BLEURT-20	4	0.686	2	0.879	8	0.486	4	0.696	6	0.681	2	0.888	10	0.484
MetaMetrics-MT-QE*	5	0.685	2	0.859	7	0.497	4	0.710	2	0.686	5	0.839	6	0.516
bright-qe*	5	0.682	3	0.817	7	0.500	2	0.794	**1**	**0.689**	5	0.806	10	0.484
BLCOM_1	6	0.664	3	0.842	11	0.455	4	0.679	6	0.681	4	0.840	9	0.488
sentinel-cand-mqm*	7	0.649	3	0.820	5	0.517	2	0.786	4	0.683	9	0.609	10	0.481
PrismRefMedium	7	0.646	4	0.776	15	0.434	4	0.651	8	0.680	3	0.872	12	0.462
PrismRefSmall	7	0.643	4	0.774	15	0.433	5	0.635	8	0.680	3	0.874	13	0.457
CometKiwi*	7	0.640	5	0.731	10	0.467	4	0.695	4	0.684	6	0.775	9	0.490
damonmonli	7	0.635	5	0.695	13	0.443	5	0.607	6	0.682	**1**	**0.912**	11	0.472
YiSi-1	8	0.630	4	0.758	14	0.436	5	0.610	7	0.681	5	0.836	13	0.458
monmonli	8	0.624	5	0.681	14	0.437	5	0.583	7	0.681	2	0.891	11	0.470
BERTScore	9	0.617	4	0.749	15	0.435	5	0.585	6	0.682	6	0.798	14	0.451
MEE4	9	0.609	5	0.731	14	0.437	7	0.498	4	0.683	3	0.856	15	0.446
chrF	10	0.607	4	0.751	17	0.431	5	0.579	9	0.680	7	0.765	18	0.436
chrfS	10	0.606	4	0.742	15	0.434	6	0.549	6	0.682	6	0.788	16	0.444
spBLEU	11	0.593	4	0.741	19	0.431	6	0.524	8	0.680	8	0.745	18	0.436
BLEU	11	0.589	4	0.736	18	0.431	7	0.513	9	0.680	8	0.739	19	0.435
BLCOM	12	0.537	6	0.619	16	0.433	3	0.730	8	0.680	10	0.325	19	0.435
sentinel-ref-mqm	12	0.523	6	0.495	20	0.429	6	0.514	9	0.680	9	0.583	19	0.435
sentinel-src-mqm*	12	0.522	6	0.496	20	0.429	7	0.512	9	0.680	9	0.581	19	0.435
*XLsimDA**	12	0.514	6	0.614	12	0.450	8	0.357	7	0.681	9	0.548	17	0.438
XLsimMqm*	12	0.514	6	0.614	12	0.450	8	0.357	7	0.681	9	0.547	17	0.438

Table 13: Soft pairwise accuracy (SPA) and acc^*_{eq} results for all metrics for main language pairs. See §5 for descriptions of SPA and acc^*_{eq}. Rows are sorted by the overall average correlation across all 6 tasks (leftmost column). Starred metrics are reference-free, underlined metrics are baselines, and italicized metrics are contrastive submissions.

Metric	avg corr	p-values																									
MetaMetrics-MT	1 0.725	.	19	07	01	01	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
MetricX-24-Hybrid	1 0.721	.	.	31	01	01	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
XCOMET	1 0.719	.	.	.	15	10	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
MetricX-24-Hybrid-QE*	2 0.714	.	.	.	.	36	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
gemba_esa*	2 0.711	.	.	.	.	.	01	00	01	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
XCOMET-QE*	3 0.695	.	.	.	.	.	.	22	14	14	02	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
COMET-22	3 0.688	.	.	.	.	.	.	.	20	34	20	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
BLEURT-20	3 0.686	.	.	.	.	.	.	.	.	43	28	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
MetaMetrics-MT-QE*	3 0.684	.	.	.	.	.	.	.	.	.	34	02	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
bright-qe*	4 0.681	.	.	.	.	.	.	.	.	.	.	06	00	00	00	00	00	00	00	00	00	00	00	00	00	00	00
BLCOM_1	4 0.664	.	.	.	.	.	.	.	.	.	.	.	04	02	00	00	01	00	00	00	00	00	00	00	00	00	00
sentinel-cand-mqm*	5 0.650	.	.	.	.	.	.	.	.	.	.	.	.	41	25	21	13	06	01	00	00	00	00	00	00	00	00
PrismRefMedium	5 0.646	.	.	.	.	.	.	.	.	.	.	.	.	.	11	35	19	01	00	00	00	00	00	00	00	00	00
PrismRefSmall	5 0.642	.	.	.	.	.	.	.	.	.	.	.	.	.	.	43	30	03	00	00	00	00	00	00	00	00	00
CometKiwi*	5 0.640	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	33	17	03	00	01	00	00	00	00	00	00
damonmonli	5 0.635	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	34	06	01	02	01	00	00	00	00	00
YiSi-1	6 0.630	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	01	00	00	00	00	00	00	00	00
BERTScore	7 0.617	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	14	04	03	00	00	00	00	00
MEE4	7 0.609	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	41	26	00	01	00	00	00
chrF	8 0.608	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	36	00	00	00	00	00
chrfS	8 0.606	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	00	01	00	00	00
spBLEU	9 0.593	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	25	00	00	00
BLEU	9 0.589	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	00	00	00
XLsimMqm*	10 0.515	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	45	49
sentinel-src-mqm*	10 0.513	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	53
sentinel-ref-mqm	10 0.513	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Table 14: 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 14 contains the results of pairwise comparisons for the results in Table 1.

C Correlations with WMT ESA for all metrics

Table 15 shows the correlations of the metrics to the ESA scores (see Section 8 for which those scores are available). The overall ranking is sorted by the average correlation, which is the average over all tasks across all language pairs. Metrics that did not participate in all tasks do not have an average correlation, and are displayed at the end of the table.

The system-level ESA scores that were used to calculate SPA here differ slightly from those in the General MT Shared Task. Namely, the General Task calculates scores by macro-averaging over domains (each domain receives equal weight), whereas we perform a standard micro-average (each segment receives equal weight).

Metric	avg-corr	cs-uk sys SPA task1	cs-uk seg acc^*_{eq} task2	en-cs sys SPA task3	en-cs seg acc^*_{eq} task4	en-es sys SPA task5	en-es seg acc^*_{eq} task6	en-hi sys SPA task7	en-hi seg acc^*_{eq} task8	en-is sys SPA task9	en-is seg acc^*_{eq} task10	en-ja sys SPA task11	en-ja seg acc^*_{eq} task12	en-ru sys SPA task13	en-ru seg acc^*_{eq} task14	en-uk sys SPA task15	en-uk seg acc^*_{eq} task16	en-zh sys SPA task17	en-zh seg acc^*_{eq} task18
MetricX-24	**1 0.708**	3 0.890	**1 0.482**	**1 0.896**	**1 0.585**	2 0.834	**1 0.503**	**1 0.938**	2 0.567	3 0.855	**1 0.670**	3 0.791	**1 0.558**	**1 0.932**	**1 0.537**	**1 0.872**	2 0.447	2 0.826	**1 0.569**
MetricX-24-Hybrid	**1 0.706**	3 0.884	**1 0.483**	2 0.886	2 0.582	**1 0.846**	2 0.496	**1 0.953**	**1 0.571**	3 0.847	2 0.661	3 0.793	**1 0.557**	2 0.921	**1 0.536**	**1 0.880**	3 0.443	2 0.808	**1 0.568**
metametrics_mt_mqm_kendall	2 0.703	5 0.819	**1 0.483**	2 0.886	3 0.575	**1 0.860**	**1 0.502**	**1 0.930**	2 0.564	4 0.846	2 0.664	3 0.787	3 0.549	**1 0.939**	**1 0.536**	2 0.850	2 0.449	**1 0.853**	2 0.564
metametrics_mt_mqm_same_source_targ	2 0.702	4 0.820	**1 0.483**	**1 0.887**	3 0.575	**1 0.858**	**1 0.502**	**1 0.928**	2 0.564	4 0.846	2 0.664	3 0.789	3 0.549	2 0.928	**1 0.536**	2 0.845	2 0.449	**1 0.854**	2 0.564
MetaMetrics-MT	3 0.702	4 0.821	**1 0.483**	2 0.885	3 0.575	**1 0.860**	**1 0.502**	**1 0.927**	2 0.564	4 0.846	2 0.664	3 0.786	3 0.549	2 0.926	**1 0.536**	2 0.849	2 0.449	**1 0.852**	2 0.564
BLEURT-20	3 0.701	**1 0.960**	3 0.471	**1 0.915**	5 0.563	3 0.793	3 0.491	2 0.906	4 0.556	2 0.888	6 0.616	2 0.824	4 0.543	**1 0.944**	4 0.520	3 0.797	5 0.434	**1 0.846**	5 0.550
XCOMET	3 0.701	6 0.782	3 0.472	**1 0.901**	3 0.572	**1 0.856**	4 0.483	2 0.918	2 0.567	3 0.866	2 0.663	4 0.778	2 0.550	**1 0.931**	2 0.531	**1 0.875**	**1 0.455**	**1 0.847**	**1 0.568**
COMET-22	3 0.700	2 0.918	2 0.477	3 0.862	4 0.566	**1 0.870**	2 0.498	**1 0.937**	4 0.552	2 0.896	3 0.650	6 0.710	3 0.548	2 0.916	3 0.528	3 0.817	3 0.441	**1 0.856**	**1 0.566**
MetricX-24-Hybrid-QE*	4 0.690	6 0.790	4 0.463	2 0.875	4 0.568	**1 0.844**	5 0.479	**1 0.934**	**1 0.568**	4 0.835	5 0.637	3 0.787	3 0.550	2 0.914	3 0.526	**1 0.872**	4 0.439	3 0.778	3 0.558
*MetricX-24-QE**	4 0.688	5 0.818	5 0.459	2 0.873	3 0.572	**1 0.838**	4 0.481	2 0.904	**1 0.569**	4 0.841	4 0.643	5 0.756	2 0.553	2 0.908	3 0.527	2 0.849	2 0.450	3 0.789	2 0.561
XCOMET-QE*	4 0.686	8 0.697	6 0.453	2 0.875	6 0.560	**1 0.859**	6 0.471	**1 0.920**	3 0.558	4 0.838	4 0.643	5 0.761	5 0.537	2 0.926	5 0.513	**1 0.884**	2 0.448	**1 0.842**	3 0.557
*CometKiwi-XXL**	5 0.681	6 0.774	6 0.453	3 0.851	5 0.562	3 0.777	6 0.473	2 0.907	3 0.561	4 0.832	3 0.645	3 0.786	5 0.536	3 0.887	5 0.514	**1 0.881**	2 0.450	2 0.809	3 0.560
YiSi-1	5 0.677	2 0.915	4 0.460	5 0.781	8 0.542	2 0.824	3 0.491	3 0.872	5 0.530	**1 0.923**	6 0.621	2 0.856	4 0.542	3 0.836	7 0.498	4 0.755	6 0.421	3 0.779	7 0.533
PrismRefSmall	5 0.676	3 0.904	3 0.471	4 0.804	7 0.548	2 0.816	6 0.473	3 0.835	4 0.556	**1 0.956**	7 0.614	2 0.846	3 0.545	3 0.841	6 0.509	4 0.731	8 0.408	3 0.759	4 0.555
PrismRefMedium	6 0.673	2 0.924	3 0.472	4 0.797	7 0.549	2 0.804	6 0.471	3 0.842	3 0.561	2 0.920	6 0.617	3 0.820	4 0.542	4 0.827	6 0.507	4 0.735	7 0.414	3 0.763	4 0.554
chrfS	7 0.666	3 0.897	5 0.458	6 0.762	9 0.534	3 0.785	4 0.481	3 0.851	5 0.528	3 0.884	7 0.612	**1 0.895**	4 0.541	3 0.832	8 0.493	5 0.721	7 0.416	3 0.767	8 0.529
chrF	8 0.662	3 0.903	6 0.454	5 0.781	10 0.530	4 0.750	5 0.479	3 0.852	6 0.523	2 0.902	7 0.615	2 0.865	6 0.528	3 0.858	9 0.486	4 0.741	8 0.412	4 0.713	9 0.522
BERTScore	9 0.656	4 0.826	6 0.455	6 0.742	10 0.528	**1 0.837**	4 0.480	4 0.820	6 0.524	3 0.868	9 0.591	**1 0.869**	6 0.531	4 0.810	9 0.485	5 0.716	7 0.413	3 0.773	8 0.531
spBLEU	9 0.652	2 0.915	9 0.440	5 0.767	11 0.527	4 0.732	6 0.473	3 0.838	7 0.515	2 0.906	8 0.601	2 0.842	7 0.516	3 0.833	11 0.473	4 0.736	9 0.408	4 0.701	10 0.518
sentinel-cand-mqm*	9 0.649	7 0.741	8 0.446	4 0.814	8 0.540	3 0.753	8 0.460	4 0.822	8 0.510	4 0.831	8 0.607	7 0.665	8 0.509	2 0.922	6 0.506	**1 0.859**	5 0.434	4 0.729	7 0.536
CometKiwi*	10 0.641	8 0.683	8 0.447	5 0.784	7 0.545	3 0.750	6 0.470	2 0.908	3 0.558	5 0.808	9 0.590	5 0.719	6 0.530	5 0.702	8 0.490	4 0.737	5 0.433	**1 0.832**	3 0.556
BLEU	10 0.637	3 0.893	11 0.415	6 0.743	13 0.513	5 0.704	7 0.469	4 0.823	9 0.500	2 0.891	10 0.567	2 0.846	7 0.514	4 0.807	12 0.460	5 0.713	10 0.402	4 0.698	10 0.517
damonmonli	10 0.633	3 0.892	6 0.455	6 0.754	14 0.508	5 0.695	8 0.463	2 0.878	7 0.518	5 0.779	11 0.550	2 0.835	7 0.519	6 0.644	10 0.478	5 0.683	9 0.404	**1 0.826**	10 0.517
*metametrics_mt_mqm_qe_same_source_t**	11 0.630	9 0.596	10 0.429	4 0.817	12 0.521	3 0.761	9 0.458	3 0.828	5 0.529	3 0.847	8 0.603	6 0.688	8 0.509	4 0.802	9 0.483	5 0.690	6 0.422	2 0.818	5 0.545
MetaMetrics-MT-QE*	11 0.630	9 0.594	10 0.429	4 0.817	12 0.521	3 0.762	9 0.458	3 0.826	5 0.529	3 0.847	8 0.603	6 0.687	8 0.509	4 0.801	9 0.483	5 0.690	6 0.422	2 0.821	6 0.545
monmonli	12 0.609	4 0.848	7 0.451	7 0.687	15 0.505	6 0.621	9 0.454	3 0.841	7 0.517	6 0.710	12 0.535	3 0.807	7 0.515	6 0.606	11 0.469	5 0.682	10 0.399	2 0.807	11 0.508
gemba_esa*	12 0.601	6 0.780	12 0.332	3 0.861	17 0.409	**1 0.838**	13 0.367	3 0.842	11 0.388	4 0.827	14 0.463	5 0.752	10 0.370	5 0.751	14 0.372	2 0.847	6 0.422	**1 0.833**	13 0.355
*XLsimDA**	13 0.496	10 0.477	11 0.413	7 0.589	16 0.480	6 0.547	11 0.429	5 0.707	10 0.481	7 0.523	13 0.493	7 0.644	9 0.486	7 0.475	13 0.444	6 0.464	11 0.382	6 0.429	12 0.471
XLsimMqm*	13 0.496	10 0.475	11 0.413	7 0.587	16 0.480	6 0.543	11 0.429	5 0.709	10 0.481	7 0.521	13 0.493	7 0.645	9 0.486	7 0.478	13 0.444	6 0.462	11 0.382	6 0.428	12 0.471
sentinel-ref-mqm	14 0.330	9 0.547	13 0.176	8 0.466	18 0.070	5 0.656	14 0.170	5 0.745	12 0.046	7 0.358	15 0.059	8 0.419	11 0.034	5 0.739	15 0.137	6 0.402	12 0.278	5 0.617	14 0.025
sentinel-src-mqm*	14 0.330	9 0.542	13 0.176	8 0.468	18 0.070	5 0.656	14 0.170	5 0.739	12 0.046	7 0.361	15 0.059	8 0.420	11 0.034	5 0.737	15 0.137	6 0.404	12 0.278	5 0.617	14 0.025
BLCOM	– –	– –	– –	– –	– –	6 0.537	10 0.445	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –
BLCOM_1	– –	– –	– –	– –	– –	**1 0.867**	2 0.497	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –
MEE4	– –	– –	– –	– –	– –	2 0.800	4 0.480	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –
bright-qe*	– –	– –	– –	– –	– –	3 0.761	12 0.396	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –	– –

Table 15: Correlations of metrics to the ESA annotations that were collected as part of the General MT Shared Task. The metrics are sorted by the average correlation across all of the correlations and language pairs. Metrics in italics are contrastive submissions and underlined metrics are baselines. QE metrics are marked by an asterisk.

D Additional figures

Figures 6-10 show the (log) p-value of two-sided paired t-test on the MQM scores against the score difference of each metric for each system pair in each language pair. Figures 11-15 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 language pair.

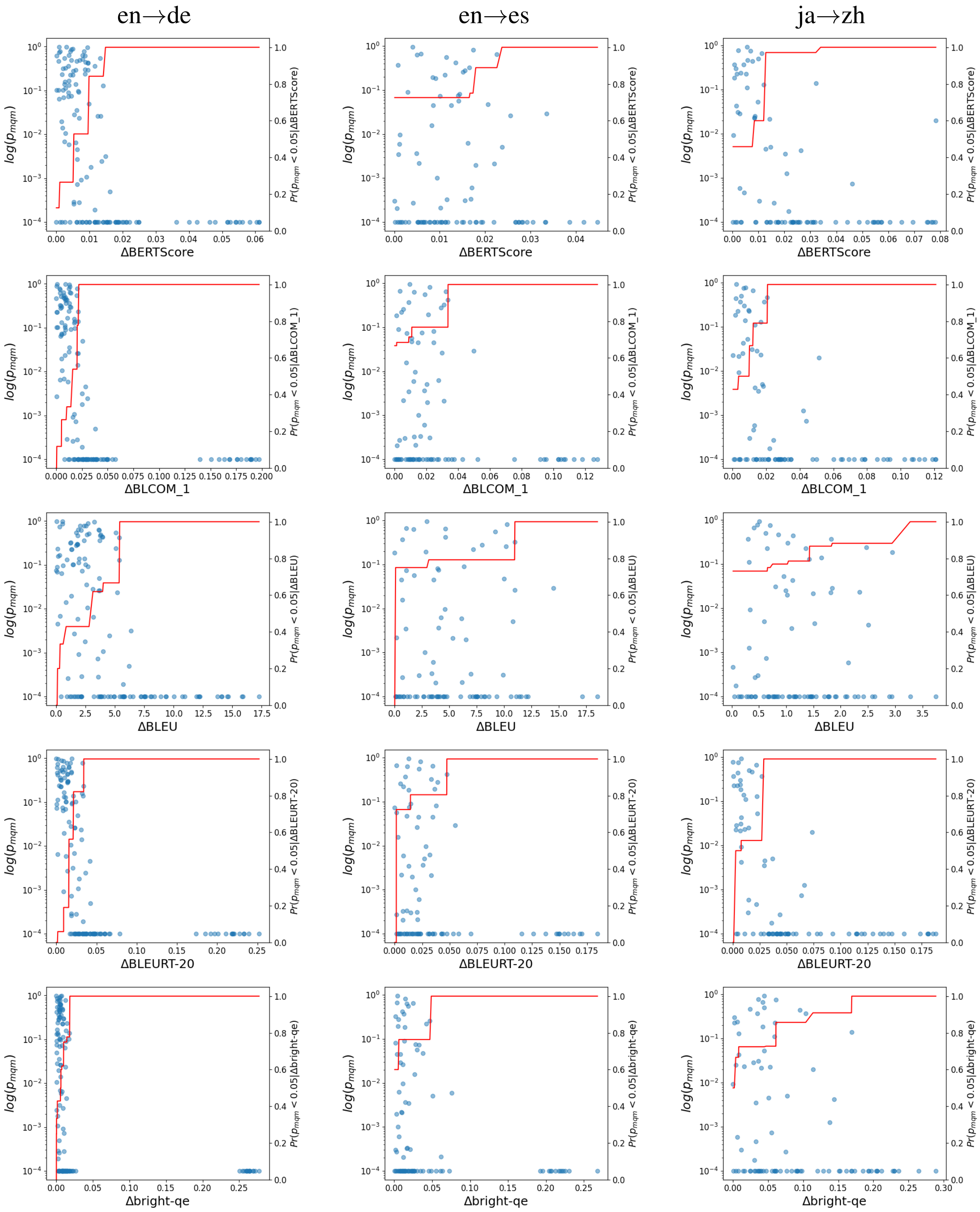

Figure 6: Log p-value of two-sided paired t-test on MQM scores (p_{mqm}) against the score difference of each metric (top to bottom: BERTSCORE, BLCOM_1, BLEU, BLEURT-20, BRIGHT-QE) for each system pair in each language pair (left to right: en→de, en→es, ja→zh). The red line is the isotonic regression 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.

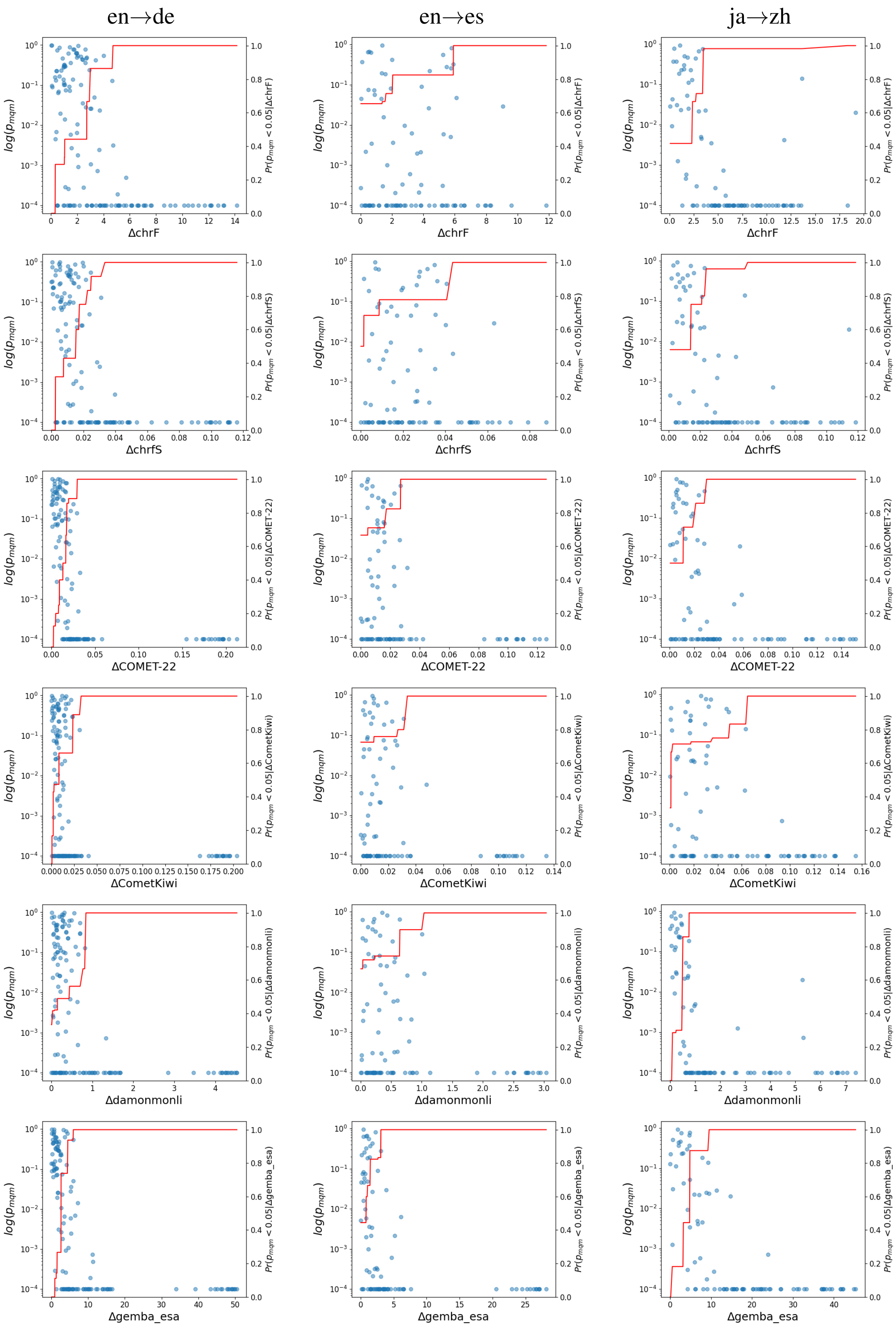

Figure 7: Log p-value of two-sided paired t-test on MQM scores (p_{mqm}) against the score difference of each metric (top to bottom: CHRF, CHRFS, COMET-22, COMETKIWI, DAMONMONLI, GEMBA_ESA) for each system pair in each language pair (left to right: en→de, en→es, ja→zh). The red line is the isotonic regression 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.

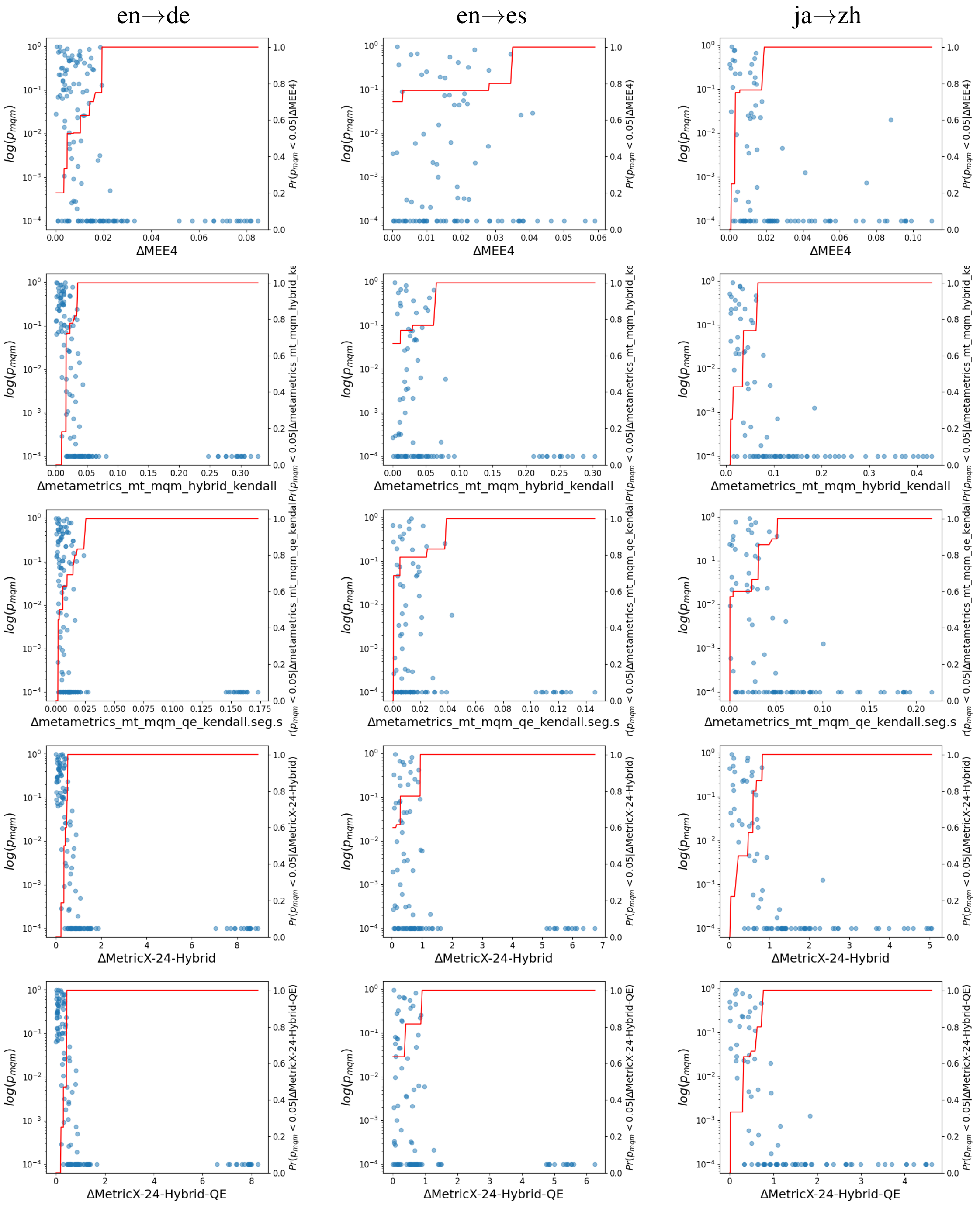

Figure 8: Log p-value of two-sided paired t-test on MQM scores (p_{mqm}) against the score difference of each metric (top to bottom: MEE4, METAMETRICS_MT_MQM_HYBRID_KENDALL, METAMETRICS_MT_MQM_QE_KENDALL.SEG.S, METRICX-24-HYBRID, METRICX-24-HYBRID-QE) for each system pair in eachlanguage pair (left to right: en→de, en→es, ja→zh). 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.

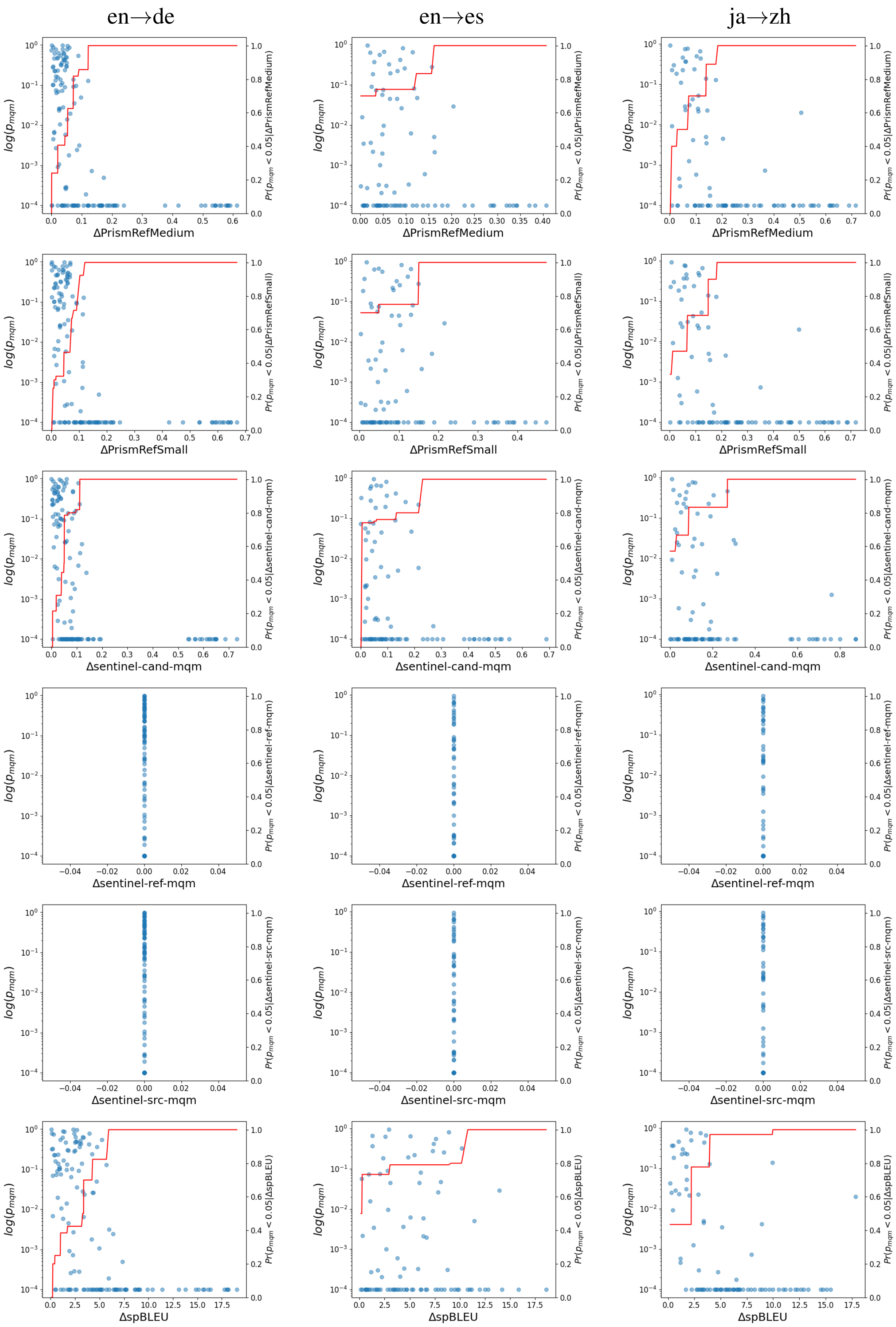

Figure 9: Log p-value of two-sided paired t-test on MQM scores (p_{mqm}) against the score difference of each metric (top to bottom: PRISMREFMEDIUM, PRISMREFSMALL, SENTINEL-CAND-MQM, SENTINEL-REF-MQM, SENTINEL-SRC-MQM, SPBLEU) for each system pair in eachlanguage pair (left to right: en→de, en→es, ja→zh). 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.

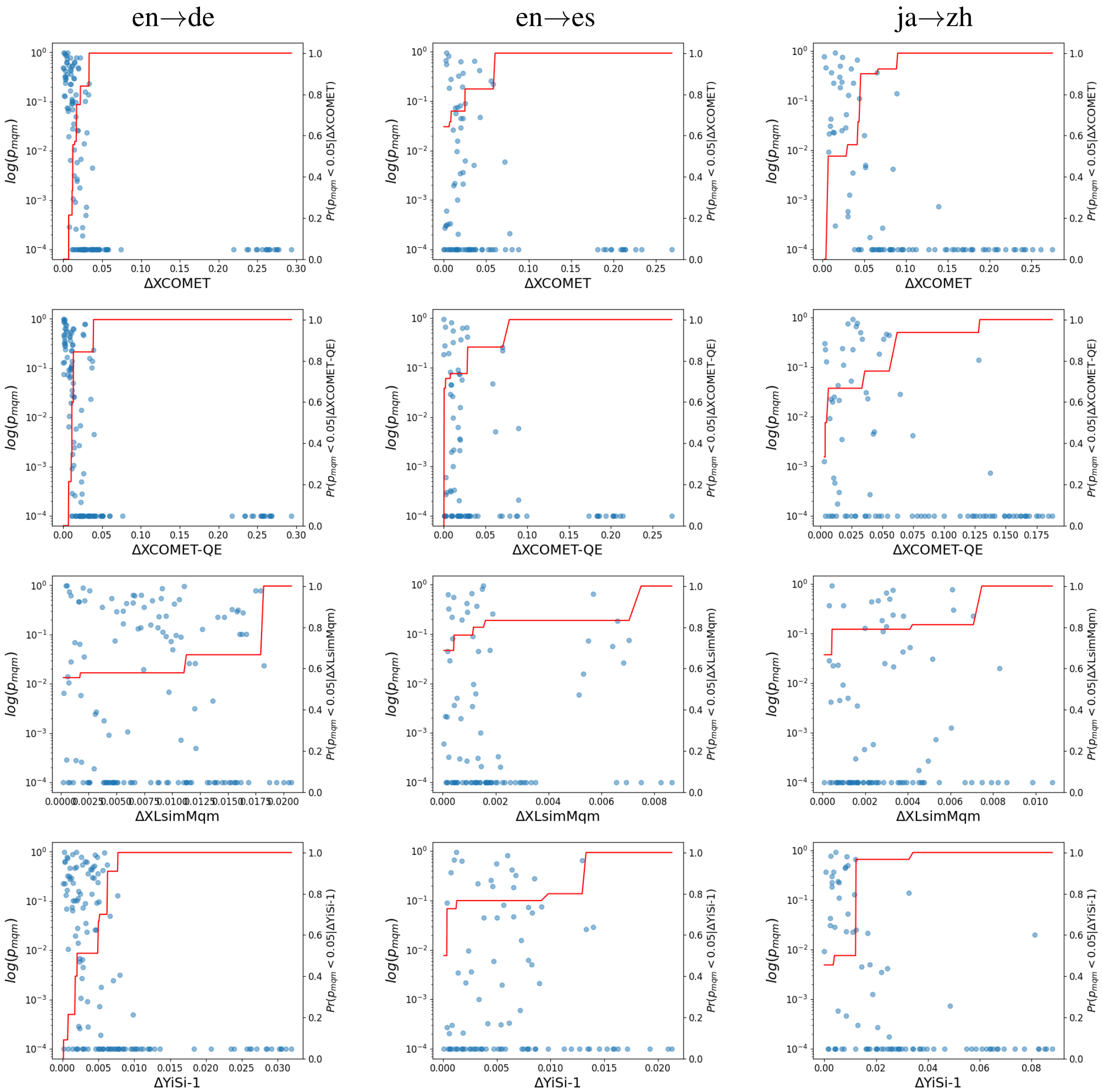

Figure 10: Log p-value of two-sided paired t-test on MQM scores (p_{mqm}) against the score difference of each metric (top to bottom: XCOMET, XCOMET-QE. XLsimMqm, YiSi-1) for each system pair in eachlanguage pair (left to right: en→de, en→es, ja→zh). The red line is the isotonic regression fit to all 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.

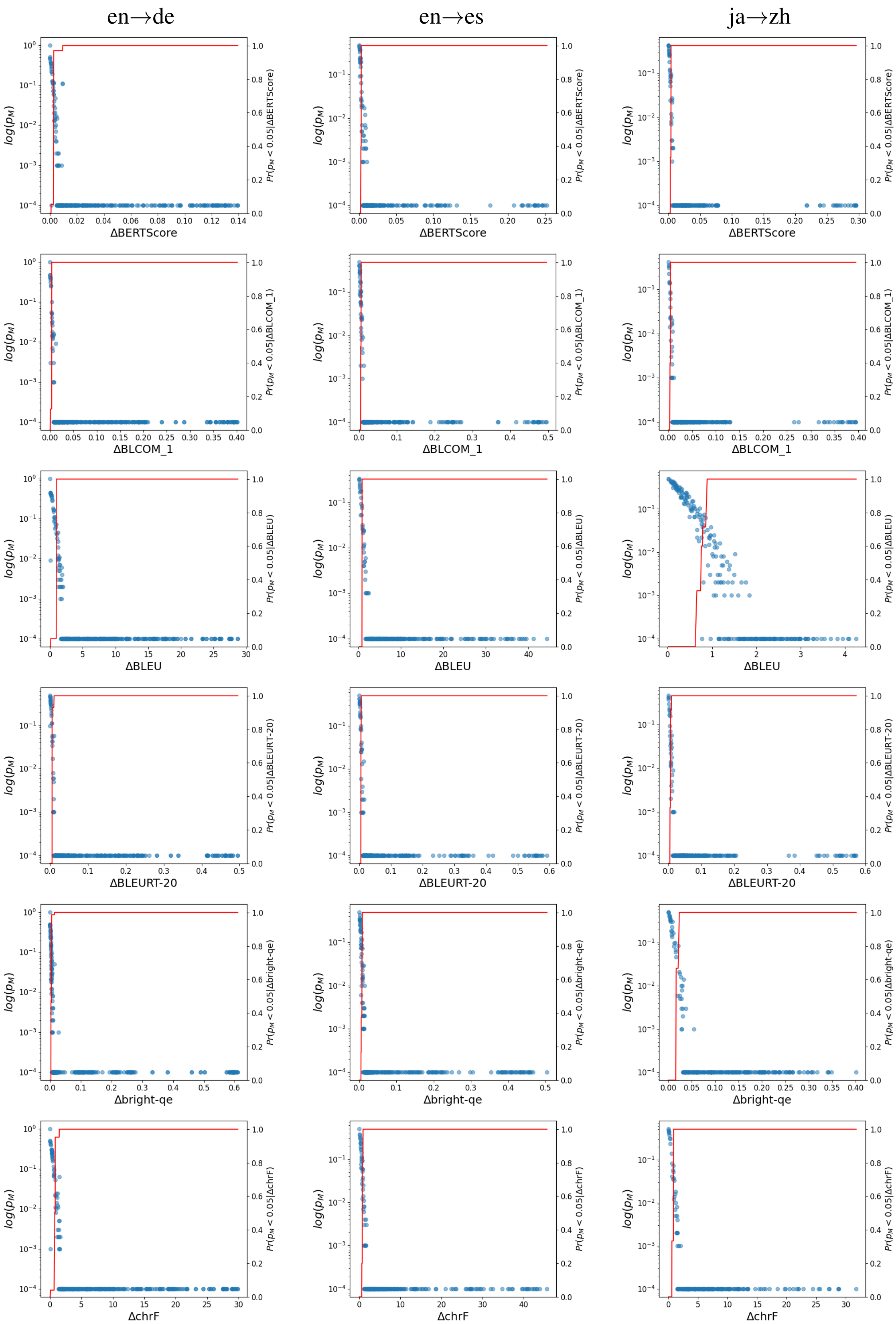

Figure 11: Log p-value of significance test with bootstrap resampling (p_M) on system-level metric scores against each metric (top to bottom: BERTSCORE, BLCOM_1, BLEU, BLEURT-20, BRIGHT-QE, CHRF) score difference for each system pair in each language pair (left to right: en→de, en→es, ja→zh). The red 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.

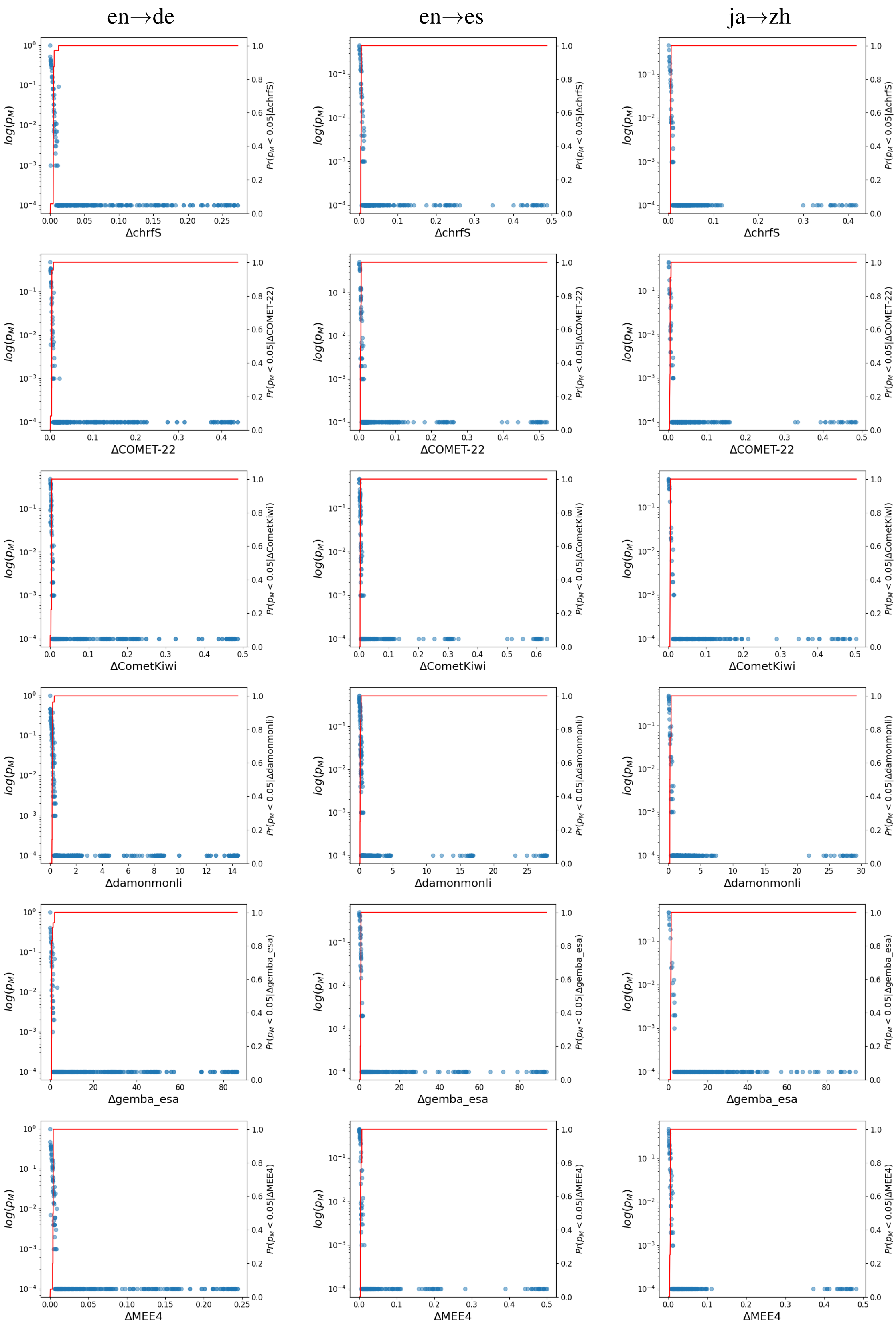

Figure 12: Log p-value of significance test with bootstrap resampling (p_M) on system-level metric scores against each metric (top to bottom: CHRFS, COMET-22, COMETKIWI, DAMONMONLI, GEMBA_ESA, MEE4) score difference for each system pair in each language pair (left to right: en→de, en→es, ja→zh). The 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.

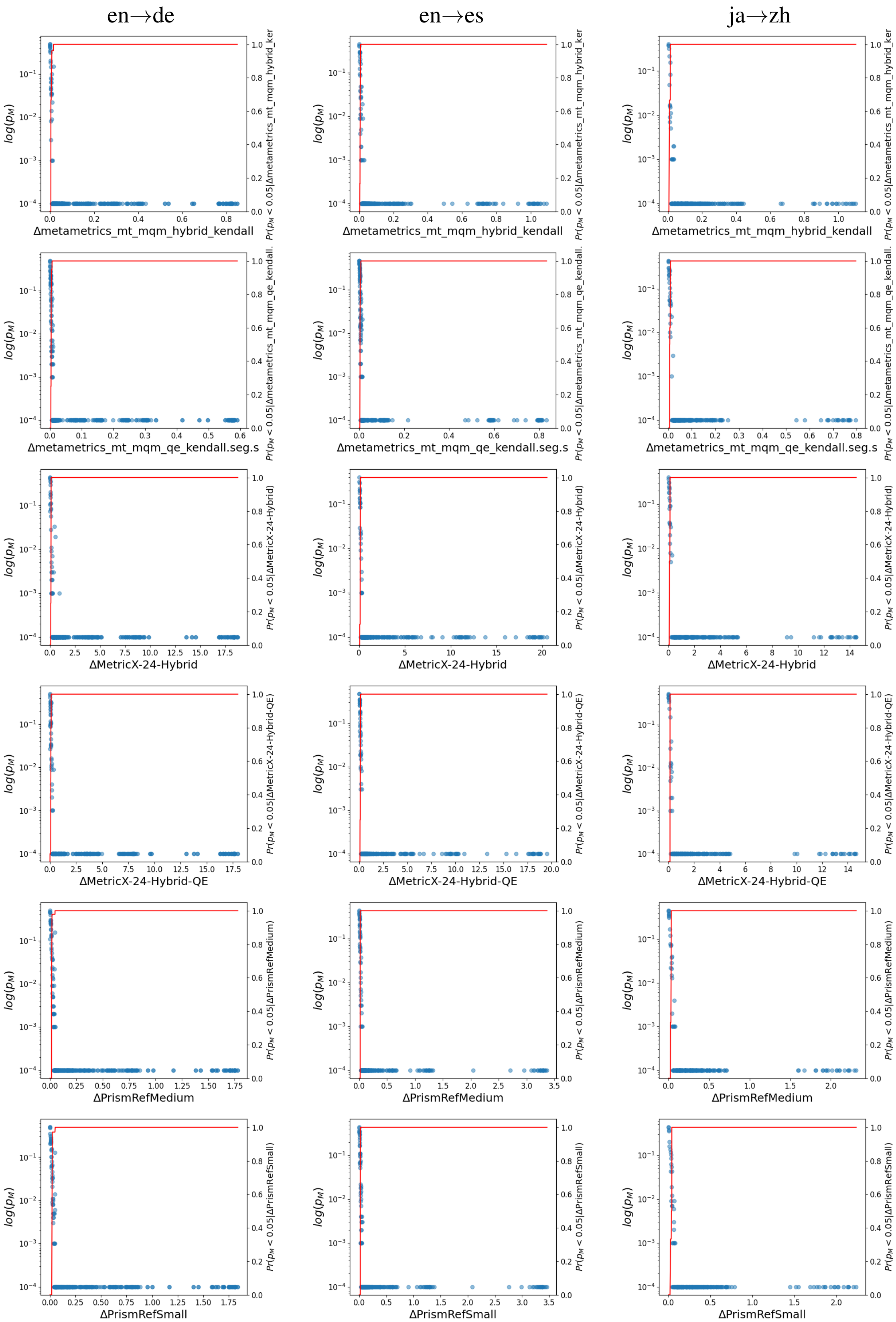

Figure 13: Log p-value of significance test with bootstrap resampling (p_M) on system-level metric scores against each metric (top to bottom: METAMETRICS_MT_MQM_HYBRID_KENDALL, METAMETRICS_MT_MQM_QE_KENDALL.SEG.S, METRICX-24-HYBRID, METRICX-24-HYBRID-QE, PRISMREFMEDIUM, PRISMREFSMALL) score difference for each system pair in each language pair (left to right: en→de, en→es, ja→zh). 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.

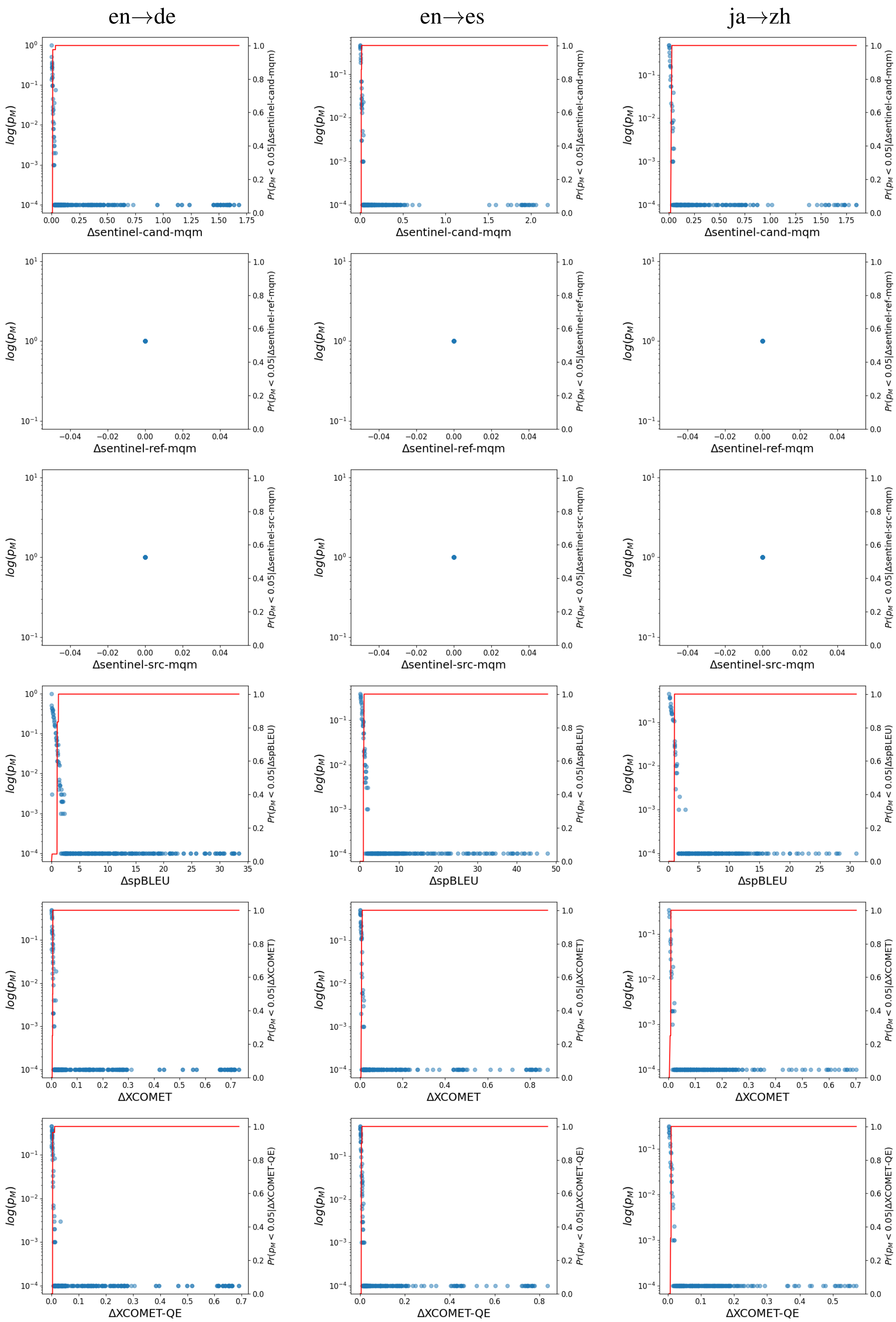

Figure 14: Log p-value of significance test with bootstrap resampling (p_M) on system-level metric scores against each metric (top to bottom: SENTINEL-CAND-MQM, SENTINEL-REF-MQM, SENTINEL-SRC-MQM, SPBLEU, XCOMET, XCOMET-QE) score difference for each system pair in each language pair (left to right: en→de, en→es, ja→zh). 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.

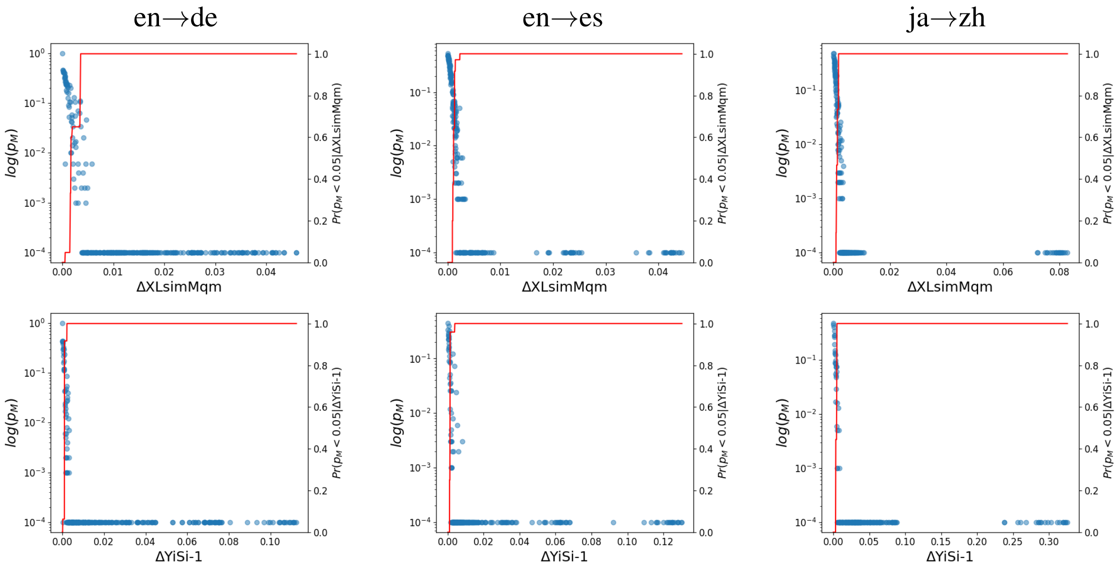

Figure 15: Log p-value of significance test with bootstrap resampling (p_M) on system-level metric scores against each metric (top to bottom: XLSIMMQM, YISI-1) score difference for each system pair in each language pair (left to right: en→de, en→es, ja→zh). The red line is the isotonic regression fit 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.