Has Machine Translation Evaluation Achieved Human Parity? The Human Reference and the Limits of Progress

Lorenzo Proietti* **Stefano Perrella*** **Roberto Navigli**
Sapienza NLP Group, Sapienza University of Rome
{lproietti, perrella, navigli}@diag.uniroma1.it

Abstract

In Machine Translation (MT) evaluation, metric performance is assessed based on agreement with human judgments. In recent years, automatic metrics have demonstrated increasingly high levels of agreement with humans. To gain a clearer understanding of metric performance and establish an upper bound, we incorporate human baselines in the MT meta-evaluation, that is, the assessment of MT metrics' capabilities. Our results show that human annotators are not consistently superior to automatic metrics, with state-of-the-art metrics often ranking on par with or higher than human baselines. Despite these findings suggesting human parity, we discuss several reasons for caution. Finally, we explore the broader implications of our results for the research field, asking: Can we still reliably measure improvements in MT evaluation? With this work, we aim to shed light on the limits of our ability to measure progress in the field, fostering discussion on an issue that we believe is crucial to the entire MT evaluation community.

1 Introduction and Related Work

Machine Translation (MT) evaluation is the task of assessing the quality of translated text, while MT meta-evaluation estimates the capabilities of automatic evaluation techniques, i.e., MT metrics. Historically, automatic metrics have been employed due to their low cost and fast experimentation time, whereas human evaluation is still considered the gold standard, necessary for validating automatically-derived findings. However, in recent years the MT evaluation field has seen significant advancements. Neural-based metrics have demonstrated strong correlations with human judgments, largely replacing traditional heuristic-based metrics, and becoming the de facto standard in MT evaluation (Freitag et al., 2022, 2023, 2024). More recently, LLM-based approaches to MT evaluation have emerged (Kocmi and Federmann, 2023b,a; Fernandes et al., 2023; Bavaresco et al., 2024), offering not only high correlation with human judgments but also improved interpretability. This raises the question of what is still missing in order for automatic techniques to achieve human parity, if they have not already. Indeed, unlike other Natural Language Processing tasks, MT evaluation lacks a human performance reference, making it difficult to gauge the true capabilities of MT metrics. For instance, in MT, human performance is measured by evaluating human references alongside system translations (Läubli et al., 2018; Kocmi et al., 2023, 2024a). Similarly, popular benchmarks such as HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), and MT-bench (Zheng et al., 2023) report the performance of human baselines.

Since MT metrics' performance is measured based on agreement with human annotators, we posit that agreement among the annotators themselves can serve as a reference for human performance. Previous studies have reported the Inter-Annotator Agreement (IAA) in MT evaluation: Lommel et al. (2014b) used Cohen's kappa to measure the pairwise agreement between raters; Freitag et al. (2021a) grouped raters' assessments into seven score bins before calculating pairwise agreement; and Kocmi et al. (2024b) used Kendall correlation coefficient τ_c. However, these studies employed different measures, making direct comparisons difficult, and none contextualized IAA in relation to the performance of automatic metrics. To the best of our knowledge, Perrella et al. (2024a) were the first to assess metric and human performance jointly. Specifically, they evaluated automatic metrics and human annotators within their new evaluation framework. Nonetheless, since comparing humans and metrics was not their primary focus, they included only one human anno-

*Equal contribution.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 790–813
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

	2020		2022		2023		2024
	→ DE	ZH →	→ DE	→ ZH	→ DE	ZH →	→ ES
MQM	3	3	3	3	4	3	1
ESA	✗	✗	✗	✗	2	✗	1
pSQM	3	3	✗	✗	✗	✗	✗
DA+SQM	✗	✗	1	1	1	1	✗
#Seg	681	895	583	1065	145	687	449
#Sys	9	9	10	13	12	15	12

Table 1: The four top rows indicate the number of distinct evaluators for each annotation protocol and test set. We list the studies that released these annotations in Appendix B. '2020' refers to the data released by Freitag et al. (2021a), while other years correspond to the test sets from the corresponding WMT editions. The notation → X indicates that the test set contains translations from English to X, whereas X → denotes translations from X to English. The two bottom rows present the number of source segments and automatic translations per source segment present in the intersection of annotations from all human evaluators, restricted to the segments annotated by disjoint sets of raters (§2.1).

tation protocol – i.e., Direct Assessment + Scalar Quality Metrics (Kocmi et al., 2022a) – that exhibited very poor performance, likely due to low annotation quality, rendering it ineffective as a human performance reference for MT metrics.

In this work, we address this gap by incorporating human baselines into the metric rankings from various editions of the Metrics Shared Task of the Conference on Machine Translation (WMT). By using meta-evaluation strategies from WMT 2024 we derive a single, comprehensive ranking of MT *evaluators* – both human and automatic – establishing a human performance reference for MT metrics across several test sets, translation directions, and human annotation protocols, and offering a clearer understanding of the capabilities of current MT evaluation techniques. Then, given that our results suggest that automatic metrics may have reached human parity, we critically examine this claim and discuss its implications for future research in MT evaluation.[1]

2 Preliminaries and Experimental Setup

In this section, we describe the human annotations, the annotation protocols, the test sets selected for our work, the meta-evaluation strategies employed, and the automatic metrics included.

[1]We publish the code to reproduce our results at `https://github.com/SapienzaNLP/human-parity-mt-eval`.

2.1 The Human Annotations

Each year, WMT conducts manual annotation campaigns to collect human judgments of translation quality. First, each test set t is created by drawing N_t segments from various sources. Segments may consist of individual sentences or entire paragraphs. Each source segment is then translated into the target language using M_t MT systems, producing $N_t \times M_t$ translations per test set t. Finally, human raters assess translation quality (Kocmi et al., 2023, 2024a; Freitag et al., 2023, 2024).

Given the large volume of translations, non-overlapping portions of each test set are typically assigned to different raters. Consequently, the annotated test sets used in this work combine annotations from multiple raters. For simplicity, we use the term *evaluator* to refer to any entity that produced a set of annotations covering all segments in a test set. An evaluator can be a human rater, an MT metric, an ensemble of MT metrics, or an entity that selects annotations from different raters. For example, in the test set that we dub "2020 EN→DE" (Freitag et al., 2021a), six raters provided a total of three annotations per translation, yielding three distinct evaluators.

However, this setup introduces a problem: Distinct human evaluators may be derived from non-disjoint sets of raters. If the same rater contributes to multiple evaluators, even across non-overlapping segments, it can artificially inflate their agreement and overestimate human baseline performance. To avoid this, we restrict each test set to the largest subset of segments annotated by strictly disjoint sets of raters. Returning to the 2020 EN→DE example, we aim to partition the six raters into three groups, so that the combined annotations from raters within each group form a single evaluator. Yet, two factors prevent such a simple partitioning: i) not all raters annotated every source segment, and ii) the specific rater-to-segment assignment prevents partitioning raters such that the combined annotations from each group cover all segments. Therefore, we restrict our test set to the segments that allow such a partitioning by solving the following optimization problem: *Find the largest subset of segments and a partitioning of raters into three disjoint groups such that each group cumulatively annotated the entire subset of segments*. We apply a similar procedure to each test set with annotations of this form, reporting resulting test set sizes in Table 1. Further details are provided in Appendix A.

Test set 2020	EN→DE				ZH→EN			
	SPA		acc^*_{eq}		SPA		acc^*_{eq}	
Metric	Rank	Acc.	Rank	Acc.	Rank	Acc.	Rank	Acc.
MQM-2020-2	1	96.45	1	58.86	1	88.10	1	55.70
pSQM-1	1	95.59	6	49.41	1	79.16	13	43.89
MQM-2020-3	2	90.39	2	56.84	1	92.06	2	52.80
BLEURT-0.2	2	86.81	4	50.81	2	72.59	3	50.57
pSQM-2	2	85.87	9	46.97	1	89.33	9	46.77
BLEURT-20	2	85.52	3	51.68	3	67.46	4	50.12

Test set 2022	EN→DE				EN→ZH			
	SPA		acc^*_{eq}		SPA		acc^*_{eq}	
Metric	Rank	Acc.	Rank	Acc.	Rank	Acc.	Rank	Acc.
MetricX-23-QE-XXL*	1	94.89	3	57.64	2	83.92	2	47.43
MQM-2022-2	1	94.49	6	55.55	2	80.82	3	47.05
MQM-2022-3	1	92.59	1	61.06	1	87.22	2	47.56
MetricX-23-XXL	2	92.34	2	59.27	1	87.69	1	48.43
DA+SQM	6	66.61	16	46.03	2	82.95	12	36.26

Test set 2023	EN→DE				ZH→EN			
	SPA		acc^*_{eq}		SPA		acc^*_{eq}	
Metric	Rank	Acc.	Rank	Acc.	Rank	Acc.	Rank	Acc.
GEMBA-MQM*	1	94.52	5	58.52	1	93.17	3	52.80
MQM-2023-3	1	93.51	5	58.42	1	95.54	5	51.65
MQM-2023-2	1	93.15	6	57.71	1	95.18	2	52.90
XCOMET-Ensemble	1	92.21	3	60.99	2	91.15	1	54.59
MetricX-23-QE-XXL*	1	92.12	1	62.53	3	88.30	2	53.26
DA+SQM	2	91.24	14	46.79	4	86.28	22	39.42
ESA-1	2	90.39	14	46.71	–	–	–	–
ESA-2	2	89.11	12	49.70	–	–	–	–
MQM-2023-4	2	88.93	14	46.68	–	–	–	–

Test set 2024	EN→ES			
	SPA		acc^*_{eq}	
Metric	Rank	Acc.	Rank	Acc.
CometKiwi-XXL*	1	86.12	4	67.24
gemba_esa*	1	85.72	3	67.68
ESA	2	80.12	8	63.84
metametrics_mt_mqm_hybrid_kendall	2	80.10	1	68.95
MetricX-24-Hybrid	2	79.75	1	69.20

Table 2: Results obtained by applying the WMT 2024 Meta-Evaluation strategies to the test sets illustrated in Section 2.2. The 'Acc.' column contains the Meta-Evaluation accuracy, while 'Rank' reports clusters of statistical significance computed following Freitag et al. (2024), using the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021). Starred metrics are reference-less metrics, and rows highlighted in gray are human evaluators.

2.2 Test Sets and Annotation Protocols

We estimate human performance based on the agreement among human evaluators. Specifically, we designate one human evaluator as ground truth while the others serve as human baselines. Consequently, our setup necessitates multiple human annotations for the same translations. Test sets satisfying this requirement include those released by Freitag et al. (2021a) and those from WMT editions between 2022 and 2024.

These test sets feature human annotations from at least two of the following protocols: Multidimensional Quality Metrics (MQM, Lommel et al., 2014a), Error Span Annotation (ESA, Kocmi et al., 2024b), Professional Scalar Quality Metrics (pSQM, Freitag et al., 2021a), and Direct Assessments + Scalar Quality Metrics (DA+SQM, Kocmi et al., 2022a). Our work leverages these test sets, but we restrict them to source segments that simultaneously: i) were annotated by all considered human evaluators and ii) were annotated by disjoint sets of raters (as detailed in Section 2.1). Table 1 presents statistics for the test sets employed. Additionally, we illustrate all the considered annotation protocols in Appendix B.

Following standard practice in the literature (Freitag et al., 2021a,b, 2022, 2023, 2024), we designate evaluators derived from the MQM annotations released annually at WMT as the ground truth, employing the others as human baselines. Indeed, the MQM protocol relies on experienced annotators and provides a more detailed (and more expensive) evaluation compared to other protocols. Nonetheless, in Appendix F, we also investigate the effects of selecting alternative evaluators – either MQM evaluators different from those previously used or evaluators following different protocols – as the ground truth.

2.3 The MT Meta-Evaluation

We compute metric rankings using the meta-evaluation strategies employed at the WMT 2024 Metrics Shared Task:

- **Soft Pairwise Accuracy (SPA)** estimates evaluator performance based on the ability to rank *MT systems*[2] in the same order as in the ranking derived from ground truth annotations (Thompson et al., 2024).

- **Pairwise Accuracy with Tie Calibration (acc^*_{eq})** estimates evaluator performance based on the ability to rank *translations of the same source segment* in the same order as in the ranking derived from ground truth annotations (Deutsch et al., 2023).

We describe these measures in more detail in Appendix C.

2.4 Metrics

The automatic evaluators considered – i.e., the MT metrics – are those submitted to the WMT Metrics Shared Task in the 2020, 2022, 2023, and 2024 editions. Additionally, we include several state-of-the-art metrics from recent WMT editions in rankings from previous years, provided they were not trained on the corresponding test sets. Table 3 in Appendix D lists all considered metrics.

[2]The score assigned to an MT system is the average of the scores given to its translations.

3 Results

Table 2 presents the evaluator rankings. Due to space constraints, each table includes only a subset of evaluators. A complete set of results, including all the evaluators, is provided in Appendix E.

Results vary across years and translation directions. Notably, human evaluators do not consistently rank higher than automatic metrics. Under SPA, human evaluators often share clusters of statistical significance with automatic metrics, whereas, under acc^*_{eq}, they are frequently surpassed. For example, in 2020 EN→DE, BLEURT-0.2 and BLEURT-20 fall within the same statistical significance cluster as MQM-2020-3 and pSQM-2 under SPA, with pSQM-2 ranking as low as 9th under acc^*_{eq}. Similarly, in 2022 EN→DE, MQM-2022-2 and MQM-2022-3 share the top cluster with MetricX-23-QE-XXL under SPA, with MQM-2022-2 ranking 6th under acc^*_{eq}. Finally, in 2023 and 2024, most human evaluators rank consistently below various automatic metrics under both SPA and acc^*_{eq}. Even when restricted to the human evaluators who follow the same protocol as the annotations employed as gold – i.e., MQM – they rank consistently in the top positions solely in 2020. Additionally, our findings remain valid when varying the human evaluators used as ground truth, as shown in Appendix F.

These results may indicate human-level performance in MT evaluation. Nonetheless, we argue that they are insufficient to establish equivalence between human and automatic evaluators, and elaborate our reasons in the next section.

4 Discussion

In the same spirit as Tedeschi et al. (2023), who discuss the meaning of superhuman performance in Natural Language Understanding, we outline several factors to consider before making similar claims in MT evaluation. We then discuss the broader implications of our findings, warning that measuring progress in the field may become increasingly challenging.

Meta-evaluation Certain meta-evaluation measures may be inadequate for comparing human and automatic evaluators. In particular, our results consistently rank human evaluators much lower under

acc^*_{eq} than under SPA. This discrepancy may be related to the findings of Perrella et al. (2024b), who show that acc^*_{eq} favors evaluators whose assessments fall within a continuous interval, whereas, as detailed in Appendix B, human evaluators produce discrete assessments.

Annotation quality Certain annotation campaigns might have produced low-quality annotations, either due to a lack of clarity in the annotation guidelines or to the ability of the raters involved. This is particularly concerning in the 2023 EN→DE test set, where, even if restricted to SPA, most human evaluators fall within the second cluster of statistical significance, alongside surface-level metrics such as BLEU.[3]

Benchmarks difficulty Current test sets might be too easy for the MT systems whose translations are being evaluated. Supporting this hypothesis, we observe that sentinel-cand-mqm, a metric that assesses only translation fluency, ranks on par with the human evaluator ESA under SPA, and even higher under acc^*_{eq} (Table 7). This suggests that the evaluated translations may differ only in minor fluency-related nuances. Arguably, to assess whether human parity has been truly achieved, future studies should compare metrics and humans in more demanding contexts. Indeed, previous research has shown that metrics may struggle in unseen domains (Zouhar et al., 2024) and lack sensitivity to specific translation errors such as incorrect number, gender (Karpinska et al., 2022), or word sense disambiguation (Martelli et al., 2025).

4.1 Can We Still Measure Improvements in MT Evaluation?

As discussed, we believe claiming human parity is premature without first addressing the issues outlined above. Nonetheless, with automatic metrics ranking the same as, or higher than, human evaluators in standard benchmarks, our results raise a critical concern about our ability to measure progress in MT evaluation: What does a higher or lower ranking truly mean?

If a metric ranks higher than a human evaluator using a non-MQM protocol, is the metric a better evaluator, or does it merely align more closely with the score distribution of the MQM protocol? More concerningly, if a metric ranks higher than an MQM evaluator, does this suggest superior evaluation capabilities, or does it simply reflect better alignment with the specific raters who produced the gold annotations? Indeed, Finkelstein et al. (2024) achieved an exceptionally high agreement with gold annotations by explicitly optimizing their metric to align with the raters themselves. More generally, we argue that in current benchmarks it is unclear whether a higher ranking – relative to either a human or an automatic evaluator – reflects genuine improvements in evaluation quality or merely closer alignment with a particular annotation protocol or rater style.

To ensure the reliability of meta-evaluation, future research should focus on exploring whether the gap between human and automatic evaluators can be restored. This could be pursued in several ways, including (but not limited to) selecting more challenging test sets, using test sets adversarial to MT metrics (e.g., from domains different from their training data), producing higher-quality human annotations, or designing new annotation protocols that yield stronger inter-annotator agreement. Additionally, greater resources could be allocated to human annotation campaigns – either by collecting multiple annotations per translation to reach a consensus among annotators or by increasing the number of segments in test sets, as suggested by Riley et al. (2024).

5 Conclusions

We incorporate human baselines into the metric rankings from previous editions of the WMT Metrics Shared Task. Our results show that MT metrics frequently rank higher than human evaluators, particularly when the latter follow annotation protocols different from MQM – the protocol used as the ground truth. While our findings suggest that metrics may have reached human-level performance, we recommend caution and highlight several issues the research community should address to assess whether human parity has been truly achieved. Finally, we discuss a critical concern arising from our findings: the limits of measuring progress in MT evaluation as automatic metrics approach human baselines. In this respect, we propose research directions to ensure that progress remains measurable or, at the very least, to extend the period during which it can be reliably tracked.

[3] We wish to highlight that our 2023 test set features only 145 segments annotated by all human evaluators (as reported in Table 1), which might have resulted in unreliable estimates of SPA and acc^*_{eq}.

Limitations

This study required test sets annotated by multiple human evaluators. Consequently, our analysis is limited to seven test sets including four language directions.

Moreover, assessing the agreement between various human evaluators required restricting our analysis to segments annotated by all of them. As a result, some test sets contain only a small number of segments, which might reduce the reliability of the results. To mitigate this issue, our findings are supported by statistical significance analyses.

Acknowledgements

We gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR, and the CREATIVE project (CRoss-modal understanding and gEnerATIon of Visual and tExtual content), which is funded by the MUR Progetti di Rilevante Interesse Nazionale programme (PRIN 2020). The authors acknowledge the CINECA award IsCb9_mtmit under the ISCRA initiative for the availability of high-performance computing resources.

This work was carried out while Lorenzo Proietti was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome. The authors thank Anna Bavaresco for her valuable feedback on the manuscript.

References

David Anugraha, Garry Kuwanto, Lucky Susanto, Derry Tanti Wijaya, and Genta Winata. 2024. MetaMetrics-MT: Tuning meta-metrics for machine translation via human preference calibration. In *Proceedings of the Ninth Conference on Machine Translation*, pages 459–469, Miami, Florida, USA. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *Preprint*, arXiv:2406.18403.

Rachel Bawden, Biao Zhang, Andre Tättar, and Matt Post. 2020. ParBLEU: Augmenting metrics with automatic paraphrases for the WMT'20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 887–894, Online. Association for Computational Linguistics.

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

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

Sören Dreano, Derek Molloy, and Noel Murphy. 2023a. Embed_Llama: Using LLM embeddings for the metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 738–745, Singapore. Association for Computational Linguistics.

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

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

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore. Association for Computational Linguistics.

Mara Finkelstein, Dan Deutsch, Parker Riley, Juraj Juraska, Geza Kovacs, and Markus Freitag. 2024. From jack of all trades to master of one: Specializing llm-based autoraters to a test set. *Preprint*, arXiv:2411.15387.

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

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian

Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

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

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

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

Thamme Gowda, Tom Kocmi, and Marcin Junczys-Dowmunt. 2023. Cometoid: Distilling strong reference-based machine translation metrics into Even stronger quality estimation metrics. In *Proceedings of the Eighth Conference on Machine Translation*, pages 751–755, Singapore. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

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*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

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

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

Tom Kocmi, 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, Steinthór 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*, pages 1–46, Miami, Florida, 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, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

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

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023b. 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, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022b. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Yilun Liu, Xiaosong Qiao, Zhanglin Wu, Su Chang, Min Zhang, Yanqing Zhao, Song Peng, Shimin Tao, Hao Yang, Ying Qin, Jiaxin Guo, Minghan Wang, Yinglu Li, Peng Li, and Xiaofeng Zhao. 2022. Partial could be better than whole. HW-TSC 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 549–557, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

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

Arle Richard Lommel, Maja Popovic, and Aljoscha Burchardt. 2014b. Assessing inter-annotator agreement for translation error annotation. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation. International Conference on Language Resources and Evaluation (LREC-14), located at LREC 14, May 26-31, Reykjavik, Iceland*. LREC.

Federico Martelli, Stefano Perrella, Niccolò Campolungo, Tina Munda, Svetla Koeva, Carole Tiberius, and Roberto Navigli. 2025. Dibimt: A gold evaluation benchmark for studying lexical ambiguity in machine translation. *Computational Linguistics*, pages 1–71.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808, Florence, Italy. Association for Computational Linguistics.

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

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

Ananya Mukherjee and Manish Shrivastava. 2022a. REUSE: REference-free UnSupervised quality estimation metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 564–568, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ananya Mukherjee and Manish Shrivastava. 2022b. Unsupervised embedding-based metric for MT evaluation with improved human correlation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 558–563, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ananya Mukherjee and Manish Shrivastava. 2023. MEE4 and XLsim : IIIT HYD‘s submissions’ for WMT23 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 800–805, Singapore. 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*, pages 470–474, Miami, Florida, USA. Association for Computational Linguistics.

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

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, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024a. Beyond correlation: Interpretable evaluation of machine translation metrics. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20689–20714, Miami, Florida, USA. Association for Computational Linguistics.

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

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

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

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

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, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

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

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

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.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020a. 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.

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

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

Peter Stanchev, Weiyue Wang, and Hermann Ney. 2019. EED: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 514–520, Florence, Italy. Association for Computational Linguistics.

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

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491, Toronto, Canada. 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. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA. Association for Computational Linguistics.

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

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

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

Vasiliy Viskov, George Kokush, Daniil Larionov, Steffen Eger, and Alexander Panchenko. 2023. Semantically-informed regressive encoder score. In *Proceedings of the Eighth Conference on Machine Translation*, pages 815–821, Singapore. Association for Computational Linguistics.

Yu Wan, Keqin Bao, Dayiheng Liu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Wenqiang Lei, and Jun Xie. 2022a. Alibaba-translate China's submission for WMT2022 metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 586–592, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022b. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Weiyue Wang, Jan-Thorsten Peter, Hendrik Rosendahl, and Hermann Ney. 2016. CharacTer: Translation edit rate on character level. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 505–510, Berlin, Germany. Association for Computational Linguistics.

Zhanglin Wu, Yilun Liu, Min Zhang, Xiaofeng Zhao, Junhao Zhu, Ming Zhu, Xiaosong Qiao, Jingfei Zhang, Ma Miaomiao, Zhao Yanqing, Song Peng, Shimin Tao, Hao Yang, and Yanfei Jiang. 2023. Empowering a metric with LLM-assisted named entity annotation: HW-TSC's submission to the WMT23 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 822–828, Singapore. Association for Computational Linguistics.

Jin Xu, Yinuo Guo, and Junfeng Hu. 2020. Incorporate semantic structures into machine translation evaluation via UCCA. In *Proceedings of the Fifth Conference on Machine Translation*, pages 934–939, Online. Association for Computational Linguistics.

Wenda Xu, Xian Qian, Mingxuan Wang, Lei Li, and William Yang Wang. 2023. SESCORE2: Learning text generation evaluation via synthesizing realistic mistakes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5166–5183, Toronto, Canada. Association for Computational Linguistics.

Wenda Xu, Yi-Lin Tuan, Yujie Lu, Michael Saxon, Lei Li, and William Yang Wang. 2022. Not all errors are equal: Learning text generation metrics using stratified error synthesis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6559–6574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

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 Fair Extraction of Evaluators from Human Annotations

The human evaluation campaigns conducted by Freitag et al. (2021a), Freitag et al. (2023), and Riley et al. (2024) produced multiple annotations for each translation. As you can see in Table 1, there are many annotations per translation for MQM and pSQM in the test sets 2020, 2022, and 2023. As discussed in Section 2.1, these annotation campaigns distributed the annotation workload among multiple raters.

Since we derive multiple evaluators from these annotations (some used as ground truth and some as human baselines), we prevent artificially inflating their agreement by not allowing the same rater to contribute to two distinct evaluators simultaneously. For example, in the 2020 EN→DE test set, six raters provided a total of three annotations per translation. We extract three human evaluators from these annotations, using one as the ground truth and the other two as human evaluators (MQM-2020-2 and MQM-2020-3 in Table 2). To achieve this, we partition the six raters into three groups, each forming one evaluator. However, not all raters annotated the entire set of source segments, and the distribution of workload did not allow for a partition that covered all annotated segments. Therefore, to retain the maximum number of segments in our test sets, we solved the following optimization problem: *Find the largest subset of segments and a partitioning of raters into three disjoint groups such that each group cumulatively annotated the entire subset of segments.*

Formally, let us define a test set $t = \{s_1, ..., s_{N_t}\}$ as a set of N_t segments. Each segment was annotated by k out of R raters, with $\mathcal{R} = \{r_1, ..., r_R\}$ representing the set of raters. Our goal is to determine a partition $\Pi = \{\mathcal{R}_1, ..., \mathcal{R}_k\}$ of $\mathcal{R}$ and a subset $u \subseteq t$ such that u is the largest subset in which every segment has been annotated by exactly one rater from each of the k sets in the partition Π.

To solve this optimization problem, we formulate it as an Integer Linear Programming (ILP) problem and solve it using the PuLP[4] Python library. We applied this procedure to the 2020, 2022, and 2023 test sets.

B Human Annotations

We briefly illustrate how each annotation protocol considered works:

- Multidimensional Quality Metrics (MQM) requires annotators to identify error spans in the translated text, specifying error category and severity, to be selected among Neutral, Minor, Major, and Critical. A translation quality score is derived by assigning a penalty to each error span depending on severity (Lommel et al., 2014a; Freitag et al., 2021a).

- Error Span Annotation (ESA) requires annotators to identify error spans in the translated text, specify error severity, and later assign a scalar quality score from 0 to 100 to the translation (Kocmi et al., 2024b).

- Scalar Quality Metrics (SQM) requires annotators to assign a scalar quality score from 0 to 6 to the translated text. Following (Freitag et al., 2021a), we use 'pSQM' to refer to SQM conducted by professional annotators.[5]

[4] https://coin-or.github.io/pulp/.

[5] In this work, we use only annotations produced by professional annotators or translators. Therefore, we exclude cSQM and Direct Assessments (DA) – which were crowdsourced – from the 2020 test sets.

- Direct Assessments + Scalar Quality Metrics (Kocmi et al., 2022a, DA+SQM) requires raters to assign a scalar quality score from 0 to 100 to the translated text. Raters are presented with seven labeled tick marks describing translation quality levels at various score thresholds, similarly to the SQM protocol.

Here, for each set of annotations employed in this work (i.e., those reported in Table 1), we indicate the reference paper that released them:

- The MQM-based and pSQM-based annotations for the test sets 2020 EN→DE and 2020 ZH→EN have been released by Freitag et al. (2021a).
- The MQM-based annotations for the test sets 2022 EN→DE and 2022 EN→ZH have been released by Freitag et al. (2022) and Riley et al. (2024).
- The DA+SQM-based annotations for the test sets 2022 EN→DE and 2022 EN→ZH have been released by Kocmi et al. (2022a).
- Three sets of MQM-based annotations for the test sets 2023 EN→DE and ZH→EN have been released by Freitag et al. (2023).
- The ESA-based annotations and the last set of MQM-based annotations (MQM-2023-4 in Table 2) for the test set 2023 EN→DE have been released by Kocmi et al. (2024b).
- The ESA-based annotations for the test set 2024 EN→ES have been released by Kocmi et al. (2024a).
- The MQM-based annotations for the test set 2024 EN→ES have been released by Freitag et al. (2024).

C Meta-Evaluation Measures

In this section, we describe the two meta-evaluation measures used in our work, as listed in Section 2.3.

C.1 Soft Pairwise Accuracy (SPA)

Thompson et al. (2024) introduced Soft Pairwise Accuracy (SPA) as an extension of Pairwise Accuracy (Kocmi et al., 2021, PA).

Given a test set t, which consists of N_t source segments and M_t translations generated by the respective M_t MT systems (as described in Section 2.1), PA counts how often an evaluator e ranks system pairs in the same order as the ground truth g. Let a_{ij} be 1 if evaluator e ranks systems i and j in the same order as the ground truth and 0 otherwise, where $i, j \in \{0, ..., Mt\}$. Then, PA is defined as:

$$PA = \binom{N}{2}^{-1} \sum_{i=0}^{M_t} \sum_{j=i+1}^{M_t} a_{ij} \tag{1}$$

SPA extends PA by incorporating the confidence with which an evaluator and the ground truth rank two MT systems. Confidence is represented using statistical p-values. Specifically, p^e_{ij} denotes the p-value associated with the statistical hypothesis that system i is better than system j according to evaluator e, while p^g_{ij} represents the corresponding p-value for the ground truth g. SPA is then defined as follows:

$$SPA = \binom{N}{2}^{-1} \sum_{i=0}^{M_t} \sum_{j=i+1}^{M_t} 1 - |p^g_{ij} - p^e_{ij}| \tag{2}$$

Thus, SPA rewards an evaluator for expressing confidence levels similar to those of the ground truth and penalizes deviations.

C.2 Pairwise Accuracy with Tie Calibration (acc^*_{eq})

Deutsch et al. (2023) introduced acc^*_{eq} to account for tied scores in meta-evaluation. Unlike PA and SPA, acc^*_{eq} is a segment-level measure, meaning it evaluates a metric's ability to estimate the quality of individual translations rather than MT systems. Specifically, acc^*_{eq} counts how often an evaluator e ranks pairs of translations of the same source segment in the same order as the ground truth g, accounting for tied scores.

Let C be the number of translation pairs ranked in the same order by both the evaluator e and the ground truth g. Similarly, let D denote the pairs ranked in the opposite order. The terms T_e and T_g represent pairs tied only in the evaluator's scores and only in the ground truth, respectively. Lastly, T_{eg} refers to pairs tied in both the evaluator's scores and the ground truth. acc^*_{eq} is then defined as:

$$\text{acc}^*_{eq} = \frac{C + T_{eg}}{C + D + T_e + T_g + T_{eg}} \tag{3}$$

Tie Calibration Many automatic metrics produce assessments on a continuous scale, such as the real numbers in the interval $[0, 1]$. As a consequence, these metrics rarely, if ever, produce tied scores, resulting in $T_e \approx 0$ and $T_{eg} \approx 0$. The Tie

Calibration algorithm addresses this issue by estimating a threshold value ϵ_e for each evaluator e, such that two assessments e_i and e_j are considered tied if $|e_i - e_j| \leq \epsilon_e$.

D Metrics

Table 3 lists the complete set of automatic evaluators considered in this work.

E Full Rankings

Tables 4, 5, 6, and 7 present the same rankings of Table 2, but including all tested evaluators.

F Full Rankings Varying the Ground Truth

In this section, we examine how evaluator rankings vary depending on the choice of human evaluator used as ground truth. Specifically, we use the following evaluators as ground truth:

- pSQM-1 from Table 4.
- DA+SQM from Table 6.
- MQM-2023-2 from Table 6.
- MQM-2023-3 from Table 6.
- ESA from Table 7.

We exclude the ESA-1, ESA-2, and MQM-2023-4 evaluators from the 2023 EN→DE test set, as they annotated only a limited number of segments. This increases the number of segments in the 2023 EN→DE test set from 145 to 376. Therefore, for reference, we also report results on this test set using the same evaluator as in Tables 2 and 6. Results are presented in Tables 8, 9, 10, 11, 12, and 13.

As we can see, our findings remain valid when varying the evaluator selected as ground truth, with human evaluators consistently ranking the same as or lower than automatic metrics.

Metric	Reference paper	Metric	Reference paper
all-rembert-20	(Mathur et al., 2020)	metametrics_mt_mqm	(Anugraha et al., 2024)
BAQ_dyn	(Mathur et al., 2020)	metametrics_mt_mqm_qe	(Anugraha et al., 2024)
BAQ_static	(Mathur et al., 2020)	MetricX-23-QE-XXL	(Juraska et al., 2023)
BERT-base-L2	(Mathur et al., 2020)	MetricX-23-XXL	(Juraska et al., 2023)
BERT-large-L2	(Mathur et al., 2020)	MetricX-24-Hybrid	(Juraska et al., 2024)
BERTScore	(Zhang et al., 2020)	MetricX-24-Hybrid-QE	(Juraska et al., 2024)
BLCOM_1	(Freitag et al., 2024)	metricx_xxl_MQM_2020	(Freitag et al., 2022)
BLEU	(Papineni et al., 2002)	mre-score-labse-regular	(Viskov et al., 2023)
BLEURT	(Sellam et al., 2020a)	MS-COMET-22	(Kocmi et al., 2022b)
BLEURT-0.1-all	(Mathur et al., 2020)	MS-COMET-QE-22	(Kocmi et al., 2022b)
BLEURT-0.1-en	(Mathur et al., 2020)	OpenKiwi-Bert	(Kepler et al., 2019)
BLEURT-0.2	(Mathur et al., 2020)	OpenKiwi-XLMR	(Kepler et al., 2019)
BLEURT-20	(Sellam et al., 2020a)	parbleu	(Bawden et al., 2020)
bleurt-combi	(Mathur et al., 2020)	parchrf++	(Bawden et al., 2020)
BLEURT-extended	(Sellam et al., 2020b)	paresim-1	(Bawden et al., 2020)
bright-qe	(Freitag et al., 2024)	prism	(Thompson and Post, 2020a)
Calibri-COMET22	(Freitag et al., 2023)	prismRef	(Thompson and Post, 2020a,b)
Calibri-COMET22-QE	(Freitag et al., 2023)	PrismRefMedium	(Thompson and Post, 2020a,b)
CharacTER	(Wang et al., 2016)	PrismRefSmall	(Thompson and Post, 2020a,b)
chrF	(Popović, 2015)	prismSrc	(Thompson and Post, 2020a,b)
chrF++	(Popović, 2017)	Random-sysname	(Freitag et al., 2023)
chrfS	(Mukherjee and Shrivastava, 2024)	REUSE	(Mukherjee and Shrivastava, 2022a)
COMET	(Rei et al., 2020b)	sentBLEU	(Papineni et al., 2002)
COMET-20	(Rei et al., 2020a)	sentinel-cand-mqm	(Perrella et al., 2024b)
COMET-22	(Rei et al., 2022a)	sentinel-ref-mqm	(Perrella et al., 2024b)
COMET-2R	(Rei et al., 2020b)	sentinel-src-mqm	(Perrella et al., 2024b)
COMET-HTER	(Rei et al., 2020b)	SEScore	(Xu et al., 2022)
COMET-MQM	(Rei et al., 2020b)	sescoreX	(Xu et al., 2023)
COMET-QE	(Rei et al., 2021)	spBLEU	(Team et al., 2022)
COMET-Rank	(Rei et al., 2020b)	SWSS+METEOR	(Xu et al., 2020)
COMETKiwi	(Rei et al., 2022b)	TER	(Snover et al., 2006)
CometKiwi-XL	(Rei et al., 2023)	tokengram_F	(Dreano et al., 2023b)
CometKiwi-XXL	(Rei et al., 2023)	UniTE	(Wan et al., 2022b,a)
cometoid22-wmt22	(Gowda et al., 2023)	UniTE-src	(Wan et al., 2022b)
damonmonli	(Freitag et al., 2024)	XCOMET	(Guerreiro et al., 2024)
docWMT22CometDA	(Vernikos et al., 2022)	XCOMET-Ensemble	(Guerreiro et al., 2024)
docWMT22CometKiwiDA	(Vernikos et al., 2022)	XCOMET-QE	(Guerreiro et al., 2024)
eBLEU	(ElNokrashy and Kocmi, 2023)	XCOMET-QE-Ensemble	(Guerreiro et al., 2024)
EED	(Stanchev et al., 2019)	XLsim	(Mukherjee and Shrivastava, 2023)
embed_llama	(Dreano et al., 2023a)	XLsimMqm	(Mukherjee and Shrivastava, 2023)
esim	(Mathur et al., 2019)	YiSi-0	(Lo, 2019)
f200spBLEU	(Team et al., 2022)	YiSi-1	(Lo, 2019)
GEMBA-MQM	(Kocmi and Federmann, 2023a)	YiSi-2	(Lo, 2019)
gemba_esa	(Freitag et al., 2024)	Yisi-combi	(Mathur et al., 2020)
HWTSC-Teacher-Sim	(Liu et al., 2022)	yisi1-translate	(Mathur et al., 2020)
KG-BERTScore	(Wu et al., 2023)	mbr-metricx-qe	(Naskar et al., 2023)
MaTESe	(Perrella et al., 2022)	mBERT-L2	(Mathur et al., 2020)
MaTESe-QE	(Perrella et al., 2022)	MEE	(Mukherjee et al., 2020)
MEE4	(Mukherjee and Shrivastava, 2022b)		

Table 3: List of all automatic evaluators considered, i.e., MT metrics, associated with their reference papers. Metrics without dedicated papers cite the Metrics Shared Task results paper in which they appeared.

	EN→DE				ZH→EN			
	SPA		acc^*_{eq}		SPA		acc^*_{eq}	
Metric	Rank	Acc.	Rank	Acc.	Rank	Acc.	Rank	Acc.
MQM-2020-2	1	96.45	1	58.86	1	88.10	1	55.70
pSQM-1	1	95.59	6	49.41	1	79.16	13	43.89
MQM-2020-3	2	90.39	2	56.84	1	92.06	2	52.80
BLEURT-0.2	2	86.81	4	50.81	2	72.59	3	50.57
pSQM-2	2	85.87	9	46.97	1	89.33	9	46.77
BLEURT-20	2	85.52	3	51.68	3	67.46	4	50.12
pSQM-3	2	84.61	6	49.38	1	87.94	7	47.88
all-rembert-20	3	79.19	4	51.04	3	66.41	3	50.61
BLEURT-extended	3	75.55	5	50.21	3	64.00	3	50.74
COMET-MQM	4	71.39	7	48.21	4	55.43	6	48.49
BLEURT-0.1-all	4	71.38	7	48.63	2	71.04	5	49.54
COMET	4	71.09	8	47.36	4	56.01	5	49.28
COMET-QE*	4	70.59	8	47.82	3	58.37	8	47.09
COMET-HTER	5	65.71	8	47.62	4	54.79	5	49.30
mBERT-L2	5	65.03	10	45.48	4	56.49	6	48.97
COMET-2R	6	58.12	9	46.43	4	55.99	4	50.20
COMET-Rank	6	54.78	14	41.31	3	58.16	14	43.57
OpenKiwi-XLMR*	6	53.25	11	44.11	4	53.29	8	47.23
OpenKiwi-Bert*	6	52.01	16	39.98	3	59.55	11	45.13
prism	6	51.92	11	43.59	4	57.88	8	47.56
Yisi-combi	7	51.10	12	42.63	–	–	–	–
bleurt-combi	7	51.10	12	42.63	–	–	–	–
esim	7	50.72	14	41.35	4	52.90	10	46.19
chrF	7	49.86	13	42.05	5	47.70	13	44.09
EED	7	49.81	15	40.94	5	45.41	14	43.64
paresim-1	7	49.54	14	41.37	4	53.34	10	46.15
chrF++	7	48.87	13	41.99	5	48.96	12	44.27
YiSi-1	7	48.79	12	42.70	4	52.74	7	48.01
CharacTER	7	47.71	16	40.45	5	48.84	13	44.01
BLEURT-0.1-en	7	47.43	15	40.96	4	57.29	7	48.26
YiSi-0	7	46.23	17	39.78	5	46.47	14	43.60
TER	7	45.98	16	40.15	6	39.68	15	43.34
parchrf++	7	45.57	13	42.25	5	48.68	12	44.25
MEE	7	45.31	14	41.61	4	52.91	13	43.94
sentBLEU	7	44.41	15	41.07	4	50.45	15	43.37
parbleu	8	41.38	15	41.01	4	50.28	15	43.43
yisi1-translate	8	39.76	12	42.60	4	52.28	11	44.70
YiSi-2*	8	38.44	18	34.36	5	43.35	12	44.60

Table 4: 2020

	EN→DE				EN→ZH			
	SPA		acc^*_{eq}		SPA		acc^*_{eq}	
Metric	Rank	Acc.	Rank	Acc.	Rank	Acc.	Rank	Acc.
MetricX-23-QE-XXL*	1	94.89	3	57.64	2	83.92	2	47.43
MQM-2022-2	1	94.49	6	55.55	2	80.82	3	47.05
MQM-2022-3	1	92.59	1	61.06	1	87.22	2	47.56
MetricX-23-XXL	2	92.34	2	59.27	1	87.69	1	48.43
COMET-22	2	91.63	5	56.51	2	84.08	3	46.74
COMET-20	2	91.28	9	52.42	2	80.56	7	43.81
CometKiwi*	2	89.51	7	53.77	3	75.36	8	43.21
BLEURT-20	3	88.20	7	53.33	3	77.80	7	43.84
metricx_xxl_MQM_2020	3	88.10	3	57.43	1	87.04	3	46.89
COMET-QE*	3	85.51	10	51.69	3	78.33	7	43.61
MS-COMET-22	3	85.37	8	53.13	1	85.18	6	44.92
CometKiwi-XXL*	3	84.43	7	53.27	2	81.25	2	47.28
UniTE	4	82.77	4	57.03	2	83.88	5	45.86
UniTE-src*	4	81.55	6	55.00	4	65.74	7	43.53
CometKiwi-XL*	4	81.13	8	52.73	2	81.56	4	46.33
YiSi-1	4	78.91	13	48.26	4	70.72	8	43.23
MATESE	5	78.03	7	53.48	–	–	–	–
BERTScore	5	75.61	14	47.57	4	70.69	8	43.28
SEScore	5	75.16	12	50.45	–	–	–	–
MS-COMET-QE-22*	5	74.44	12	50.37	2	78.84	9	42.51
MEE4	5	74.19	15	46.81	–	–	–	–
chrF	5	73.05	16	46.38	3	72.67	10	41.87
f200spBLEU	5	71.04	15	46.84	4	71.76	10	41.85
HWTSC-Teacher-Sim*	5	69.68	13	48.10	4	68.43	11	40.53
DA+SQM	6	66.61	16	46.03	2	82.95	12	36.26
MATESE-QE*	6	65.42	11	51.06	–	–	–	–
BLEU	6	65.00	15	46.51	4	67.31	13	34.28
REUSE*	7	37.95	17	43.58	5	33.46	12	35.89

Table 5: 2022

Metric	EN→DE SPA Rank	EN→DE SPA Acc.	EN→DE acc^*_{eq} Rank	EN→DE acc^*_{eq} Acc.	ZH→EN SPA Rank	ZH→EN SPA Acc.	ZH→EN acc^*_{eq} Rank	ZH→EN acc^*_{eq} Acc.
GEMBA-MQM*	1	94.52	5	58.52	1	93.17	3	52.80
MQM-2023-3	1	93.51	5	58.42	1	95.54	5	51.65
CometKiwi-XXL*	1	93.22	5	58.46	1	92.86	6	50.94
MQM-2023-2	1	93.15	6	57.71	1	95.18	2	52.90
CometKiwi-XL*	1	93.11	6	57.38	2	92.02	6	50.71
MetricX-23-XXL	1	92.57	2	61.82	2	91.58	2	53.13
XCOMET-QE-Ensemble*	1	92.48	4	59.89	2	90.54	3	52.87
cometoid22-wmt22*	1	92.43	5	58.09	2	90.09	7	50.23
COMET	1	92.33	7	56.65	4	87.18	9	48.42
XCOMET-Ensemble	1	92.21	3	60.99	2	91.15	1	54.59
MetricX-23-QE-XXL*	1	92.12	1	62.53	3	88.30	2	53.26
Calibri-COMET22	1	92.01	10	51.26	4	87.02	16	44.57
docWMT22CometDA	1	91.76	8	54.71	4	87.41	13	46.15
sescoreX	1	91.66	8	54.76	4	85.73	13	46.39
DA+SQM	2	91.24	14	46.79	4	86.28	22	39.42
Calibri-COMET22-QE*	2	90.80	11	50.33	4	87.59	11	47.24
ESA-1	2	90.39	14	46.71	–	–	–	–
BLEURT-20	2	90.35	8	55.19	4	87.36	9	48.63
mbr-metricx-qe*	2	89.98	5	58.75	3	88.55	4	52.04
prismRef	2	89.92	11	50.72	5	82.50	14	46.06
docWMT22CometKiwiDA*	2	89.92	8	55.30	2	90.95	10	47.83
MS-COMET-QE-22*	2	89.85	8	54.55	4	87.59	10	47.81
f200spBLEU	2	89.24	11	50.54	5	81.12	18	43.33
CometKiwi*	2	89.23	6	57.74	3	89.37	5	51.74
mre-score-labse-regular	2	89.14	10	51.12	4	87.14	17	43.80
ESA-2	2	89.11	12	49.70	–	–	–	–
YiSi-1	2	88.96	9	53.15	4	85.70	12	46.68
MQM-2023-4	2	88.93	14	46.68	–	–	–	–
KG-BERTScore*	2	88.79	7	56.98	3	89.31	8	49.75
MaTESe	2	88.40	9	53.36	2	92.06	7	50.34
BLEU	2	88.02	12	50.06	6	80.92	19	43.13
BERTscore	2	87.33	11	50.88	5	84.68	15	45.79
MEE4	2	87.07	10	51.62	6	80.51	19	42.94
XLsim	2	86.58	10	51.01	6	81.00	19	42.84
tokengram_F	3	85.60	12	49.72	5	81.01	18	43.52
chrF	4	84.25	12	49.54	5	81.47	17	43.72
eBLEU	4	83.87	13	48.96	6	80.44	20	42.55
embed_llama	4	81.33	14	47.12	4	84.84	21	41.05
Random-sysname*	5	59.47	16	39.07	7	54.34	23	34.49
prismSrc*	6	30.03	15	40.89	8	35.54	22	39.28

Table 6: 2023

	EN→ES			
	SPA		acc^*_{eq}	
Metric	Rank	Acc.	Rank	Acc.
CometKiwi-XXL*	1	86.12	4	67.24
gemba_esa*	1	85.72	3	67.68
COMET-22	1	82.37	5	66.60
bright-qe*	1	81.77	4	67.39
ESA	2	80.12	8	63.84
XCOMET-QE*	2	80.10	3	67.99
metametrics_mt_mqm_hybrid_kendall	2	80.10	1	68.95
XCOMET	2	79.96	2	68.67
MetricX-24-Hybrid	2	79.75	1	69.20
BLCOM_1	2	79.17	6	65.02
MetricX-24-Hybrid-QE*	2	79.09	2	68.92
sentinel-cand-mqm*	2	78.54	5	66.39
BLEURT-20	2	75.96	7	64.48
metametrics_mt_mqm_qe_kendall.seg.s*	3	73.29	4	67.49
CometKiwi*	3	71.74	5	66.51
PrismRefMedium	3	70.93	11	61.39
PrismRefSmall	3	70.52	10	61.51
YiSi-1	3	70.51	11	61.44
BERTScore	3	67.75	11	61.41
chrF	3	66.73	13	61.05
damonmonli	3	66.37	9	62.10
chrfS	4	64.31	11	61.37
spBLEU	4	63.19	12	61.08
BLEU	5	60.67	13	61.04
MEE4	5	60.36	10	61.57
sentinel-ref-mqm	6	44.19	13	61.04
sentinel-src-mqm*	6	44.19	13	61.04
XLsimMqm*	6	39.25	12	61.11

Table 7: 2024

	ZH→EN			
	SPA		acc^*_{eq}	
Metric	Rank	Acc.	Rank	Acc.
pSQM-3	1	83.42	6	65.23
MQM-2020-2	1	80.56	6	65.22
MQM-2020-3	1	80.34	6	65.22
MQM-2020-1	1	79.55	6	65.22
pSQM-2	1	74.02	4	65.26
BERT-large-L2	1	70.19	3	65.38
COMET	1	68.58	1	65.64
SWSS+METEOR	2	67.26	4	65.26
MEE	2	67.07	6	65.22
prism	2	66.01	5	65.25
sentBLEU	2	65.70	5	65.25
parbleu	2	65.63	6	65.23
BLEURT	2	65.28	2	65.49
YiSi-1	2	64.71	6	65.22
yisi1-translate	2	64.32	6	65.23
CharacTER	2	63.99	6	65.23
all-rembert-20	2	63.48	2	65.47
BLEURT-20	2	63.10	3	65.38
paresim-1	2	62.79	4	65.30
esim	2	62.42	4	65.30
chrF++	3	62.38	6	65.23
BLEURT-0.1-en	3	62.22	2	65.47
COMET-2R	3	62.10	1	65.61
parchrf++	3	61.92	5	65.24
EED	3	60.91	6	65.23
mBERT-L2	3	60.84	2	65.43
chrF	3	60.81	4	65.25
YiSi-0	3	60.55	5	65.25
BLEURT-0.2	3	60.54	3	65.34
COMET-Rank	3	59.97	6	65.22
BLEURT-extended	3	59.90	3	65.36
BAQ_static	3	59.89	6	65.22
BLEURT-0.1-all	3	59.80	4	65.25
BERT-base-L2	3	59.57	2	65.49
COMET-HTER	3	58.88	2	65.42
COMET-MQM	3	58.27	5	65.25
BAQ_dyn	3	58.12	6	65.22
COMET-QE*	3	57.83	4	65.26
TER	3	56.12	5	65.25
OpenKiwi-Bert*	3	55.35	5	65.25
OpenKiwi-XLMR*	3	50.98	4	65.30
YiSi-2*	4	48.35	4	65.27

Table 8: The test set is 2020 ZH→EN. The evaluator selected as the ground truth follows the pSQM protocol (pSQM-1 in Table 4).

	EN→DE			
	SPA		acc^*_{eq}	
Metric	Rank	Acc.	Rank	Acc.
MetricX-23-QE-XXL*	1	95.11	3	58.41
GEMBA-MQM*	1	94.89	14	43.06
CometKiwi*	1	94.83	3	58.21
KG-BERTScore*	1	94.57	5	57.16
MQM-2023-3	1	94.24	13	46.99
docWMT22CometKiwiDA*	1	94.22	2	58.60
CometKiwi-XL*	1	94.16	2	58.80
MS-COMET-QE-22*	1	94.04	7	55.66
CometKiwi-XXL*	1	93.98	1	59.66
MetricX-23-XXL	2	93.31	3	58.06
COMET	2	92.80	3	58.42
docWMT22CometDA	2	92.45	2	58.95
mre-score-labse-regular	2	92.30	8	54.75
MQM-2023-1	2	92.10	15	42.20
Calibri-COMET22	2	92.02	3	58.17
mbr-metricx-qe*	2	91.94	3	58.26
MQM-2023-2	2	91.40	11	48.91
cometoid22-wmt22*	2	91.34	5	57.06
sescoreX	2	91.22	4	57.41
BLEURT-20	3	90.66	5	57.26
prismRef	3	89.58	10	54.16
Calibri-COMET22-QE*	3	89.55	8	55.22
YiSi-1	3	89.41	6	56.64
XLsim	3	87.93	6	56.47
XCOMET-Ensemble	4	87.77	4	57.82
XCOMET-QE-Ensemble*	4	87.34	6	56.40
eBLEU	4	87.33	10	53.91
BERTscore	4	86.83	7	56.13
f200spBLEU	4	86.82	8	54.97
MaTESe	4	86.69	16	37.35
MEE4	4	86.12	7	55.93
BLEU	5	84.45	10	53.63
tokengram_F	5	83.15	8	54.98
chrF	5	82.45	9	54.74
embed_llama	5	81.27	9	54.28
Random-sysname*	6	60.55	12	47.94
prismSrc*	7	28.59	11	48.54

Table 9: The test set is 2023 EN→DE. The evaluator selected as the ground truth follows the DA+SQM protocol (DA+SQM in Table 6). Different from Tables 2 and 6, we exclude the evaluators ESA-1, ESA-2, and MQM-2023-4, because they annotated a limited number of translations. This way, we increase the number of segments in the test set from 145 to 376.

	EN→DE			
	SPA		acc^*_{eq}	
Metric	Rank	Acc.	Rank	Acc.
MQM-2023-3	1	97.09	7	56.62
GEMBA-MQM*	1	97.09	5	59.18
CometKiwi-XXL*	1	95.96	7	56.43
CometKiwi-XL*	1	95.47	6	57.33
MQM-2023-2	1	95.20	4	60.04
docWMT22CometDA	1	94.99	10	53.86
MetricX-23-XXL	2	94.93	2	61.65
XCOMET-Ensemble	2	94.58	1	62.23
MetricX-23-QE-XXL*	2	94.52	3	61.14
COMET	2	94.39	8	55.71
XCOMET-QE-Ensemble*	2	94.14	4	59.89
docWMT22CometKiwiDA*	2	93.70	10	53.55
BLEURT-20	2	93.35	9	54.83
Calibri-COMET22-QE*	2	93.25	14	49.62
cometoid22-wmt22*	2	92.43	7	56.73
CometKiwi*	3	92.27	7	56.76
DA+SQM	3	92.10	18	43.68
KG-BERTScore*	3	92.05	9	54.40
sescoreX	3	91.99	10	53.99
YiSi-1	3	91.28	11	51.66
mbr-metricx-qe*	3	91.22	8	56.19
MS-COMET-QE-22*	3	90.70	10	53.51
prismRef	3	90.22	14	49.74
Calibri-COMET22	3	88.28	14	49.54
XLsim	4	88.01	13	50.31
mre-score-labse-regular	4	87.47	13	49.85
BERTscore	4	87.01	12	50.48
f200spBLEU	4	86.90	13	50.37
MaTESe	4	86.73	7	56.23
eBLEU	4	86.24	16	48.30
MEE4	4	86.05	12	50.81
BLEU	5	84.48	15	49.18
tokengram_F	5	83.85	14	49.51
chrF	5	83.34	14	49.44
embed_llama	5	80.09	17	45.05
Random-sysname*	6	61.26	20	38.19
prismSrc*	7	28.88	19	40.02

Table 10: The test set is 2023 EN→DE. The evaluator selected as the ground truth follows the MQM protocol (it is the evaluator selected as ground truth in Table 6). Different from Tables 2 and 6, we exclude the evaluators ESA-1, ESA-2, and MQM-2023-4, because they annotated a limited number of translations. This way, we increase the number of segments in the test set from 145 to 376.

	EN→DE			
	SPA		acc^*_{eq}	
Metric	Rank	Acc.	Rank	Acc.
BLEURT-20	1	97.12	6	60.66
MetricX-23-XXL	1	96.51	1	64.61
docWMT22CometDA	1	96.41	8	59.16
CometKiwi-XXL*	1	96.40	4	61.57
GEMBA-MQM*	1	96.19	8	58.96
CometKiwi-XL*	1	95.88	6	60.45
COMET	1	95.82	5	61.31
MQM-2023-3	1	95.50	8	59.25
MQM-2023-1	1	95.20	7	60.03
mbr-metricx-qe*	1	95.12	3	62.39
MetricX-23-QE-XXL*	2	94.99	2	63.51
Calibri-COMET22-QE*	2	94.48	15	51.95
docWMT22CometKiwiDA*	2	94.47	9	57.58
XCOMET-Ensemble	2	93.81	1	64.20
sescoreX	2	93.51	7	59.94
CometKiwi*	2	93.04	6	60.24
KG-BERTScore*	2	92.78	8	59.06
XCOMET-QE-Ensemble*	3	92.71	4	61.92
DA+SQM	3	91.40	16	48.92
MS-COMET-QE-22*	3	91.34	9	57.54
cometoid22-wmt22*	3	91.22	5	61.07
YiSi-1	3	90.47	9	57.50
mre-score-labse-regular	3	90.02	10	56.50
prismRef	3	89.92	12	55.16
f200spBLEU	3	89.09	11	55.81
XLsim	4	88.77	11	55.49
Calibri-COMET22	4	88.37	11	55.67
eBLEU	4	88.19	14	54.13
BERTscore	4	88.09	10	56.27
BLEU	4	87.09	13	54.61
MaTESe	4	86.49	13	54.19
MEE4	4	86.46	10	56.41
tokengram_F	4	86.12	11	55.66
chrF	4	85.53	11	55.50
embed_llama	5	80.95	15	51.39
Random-sysname*	6	59.64	18	42.43
prismSrc*	7	25.26	17	43.73

Table 11: The test set is 2023 EN→DE. The evaluator selected as the ground truth follows the MQM protocol (MQM-2023-2 in Table 6). Different from Tables 2 and 6, we exclude the evaluators ESA-1, ESA-2, and MQM-2023-4, because they annotated a limited number of translations. This way, we increase the number of segments in the test set from 145 to 376.

	EN→DE			
	SPA		acc^*_{eq}	
Metric	Rank	Acc.	Rank	Acc.
GEMBA-MQM*	1	97.98	7	55.71
CometKiwi-XXL*	1	97.96	4	58.36
CometKiwi-XL*	1	97.49	5	57.78
MQM-2023-1	1	97.09	6	56.46
MetricX-23-XXL	1	96.86	1	61.26
docWMT22CometDA	1	96.82	6	56.98
MetricX-23-QE-XXL*	1	96.47	2	60.55
COMET	2	96.30	4	58.63
docWMT22CometKiwiDA*	2	96.09	8	54.86
MQM-2023-2	2	95.50	3	59.25
CometKiwi*	2	94.67	5	57.83
KG-BERTScore*	2	94.45	6	56.67
sescoreX	2	94.31	5	57.58
DA+SQM	2	94.24	14	46.99
BLEURT-20	3	94.18	4	58.68
Calibri-COMET22-QE*	3	94.01	13	49.98
YiSi-1	3	93.19	7	55.64
MS-COMET-QE-22*	3	93.17	7	55.46
mbr-metricx-qe*	3	93.06	3	59.30
XCOMET-Ensemble	3	92.64	2	60.61
prismRef	3	92.58	10	53.57
cometoid22-wmt22*	3	92.55	4	58.50
XCOMET-QE-Ensemble*	3	92.14	3	59.43
Calibri-COMET22	4	90.86	10	53.18
XLsim	4	90.86	8	54.63
BERTscore	4	89.67	9	54.30
f200spBLEU	4	89.56	9	54.06
mre-score-labse-regular	4	89.14	8	54.47
MEE4	4	88.72	8	54.40
eBLEU	4	88.67	11	53.06
BLEU	5	87.13	11	52.76
tokengram_F	5	86.17	10	53.62
chrF	5	85.58	10	53.51
MaTESe	5	84.75	12	52.12
embed_llama	5	82.93	13	50.10
Random-sysname*	6	58.48	16	41.27
prismSrc*	7	28.67	15	44.52

Table 12: The test set is 2023 EN→DE. The evaluator selected as the ground truth follows the MQM protocol (MQM-2023-3 in Table 6). Different from Tables 2 and 6, we exclude the evaluators ESA-1, ESA-2, and MQM-2023-4, because they annotated a limited number of translations. This way, we increase the number of segments in the test set from 145 to 376.

	EN→ES			
	SPA		acc^*_{eq}	
Metric	Rank	Acc.	Rank	Acc.
COMET-22	1	86.90	2	53.11
BLCOM_1	1	86.12	2	53.00
XCOMET	1	84.88	3	52.35
metametrics_mt_mqm_hybrid_kendall	1	84.80	1	53.92
XCOMET-QE*	1	83.67	4	51.01
PrismRefMedium	1	83.07	5	50.62
MetricX-24-Hybrid	2	82.96	2	53.13
BLEURT-20	2	82.19	2	52.93
gemba_esa*	2	81.37	10	40.86
MetricX-24-Hybrid-QE*	2	80.93	4	51.35
PrismRefSmall	2	80.84	4	51.23
MQM-2024	2	80.13	11	34.61
sentinel-cand-mqm*	2	79.06	6	50.23
YiSi-1	2	78.76	3	51.88
BERTScore	2	78.28	4	50.88
metametrics_mt_mqm_qe_kendall.seg.s*	3	77.05	7	49.17
CometKiwi-XXL*	3	76.51	5	50.86
bright-qe*	3	75.73	9	43.46
MEE4	3	75.20	4	51.03
CometKiwi*	3	74.39	5	50.47
chrfS	3	73.90	4	51.05
chrF	4	70.94	5	50.69
spBLEU	4	70.77	6	49.81
BLEU	4	69.37	7	49.43
damonmonli	4	63.34	8	48.14
sentinel-ref-mqm	5	54.36	12	15.38
sentinel-src-mqm*	5	54.36	12	15.38
XLsimMqm*	5	39.32	9	43.02

Table 13: The test set is 2024. The evaluator selected as the ground truth follows the ESA protocol (ESA in Table 7).