Can Automatic Metrics Assess High-Quality Translations?

Anonymous ACL submission

Abstract

Automatic metrics for evaluating translation quality are typically validated by measuring how well they correlate with human assessments. However, correlation methods tend to capture only the ability of metrics to differentiate between good and bad source-translation pairs, overlooking their reliability in distinguishing alternative translations for the same source. In this paper, we confirm that this is indeed the case by showing that current metrics are insensitive to nuanced differences in translation quality. This effect is most pronounced when the quality is high and the variance among alternatives is low. Given this finding, we shift towards detecting high-quality correct translations, an important problem in practical decision-making scenarios where a binary check of correctness is prioritized over a nuanced evaluation of quality. Using the MQM framework as the gold standard, we systematically stress-test the ability of current metrics to identify translations with no errors as marked by humans. Our findings reveal that current metrics often over or underestimate translation quality, indicating significant room for improvement in machine translation evaluation.

1 Introduction

The automatic evaluation of machine or human-generated translations has gained widespread attention over the past few years. Evaluation metrics act as proxies for translation quality in the absence of human judgments, offering immediate feedback. They are widely used not only to provide quality indicators to users and translators (Béchara et al., 2021; Castilho and O'Brien, 2017; Mehandru et al., 2023a), but also to improve machine translation (MT) systems themselves (He et al., 2024; Xu et al., 2024a; Fernandes et al., 2022).

Judging whether, and to what extent, these metrics concur with human evaluation is important to ensuring their effectiveness and applicability in diverse scenarios. A recent human evaluation study by the Conference on Machine Translation (WMT) revealed that translations produced by current MT systems often achieve very high-quality scores (ranging from 80 to 90) when judged by humans on a direct assessment (DA) scale of 0 to 100 (Kocmi et al., 2023). Similarly, Deutsch et al. (2023) observe that these systems increasingly generate numerous "perfect" translations (translations with zero errors), especially for high-resource language pairs, as shown in Table 1. As MT quality advances, evaluating whether evaluation metrics accurately reflect this progress is essential. The absence of clear criteria for assessing these high-quality translations can introduce bias, leading to inconsistent assessments based on metric preferences rather than objective measures of accuracy.

LP	N	% ZERO-MQM	
WMT 2023 METRICS DATASET			
EN-DE (P)	5520	25.4%	
HE-EN	9840	50.8%	
ZH-EN	17655	19.1%	
WMT 2022 METRICS DATASET			
EN-DE	18410	51.5%	
EN-RU	19725	42.7%	
ZH-EN	26250	46.4%	
WMT 2022 CHAT DATASET			
XX-EN	4756	63.2%	
EN-XX	5901	60.2%	

Table 1: Gold MQM scores distribution in recent WMT datasets. High-quality translations are represented in shades of green (darker for MQM $= 0$ and lighter for MQM ≥ -5); red represents translations with at least one major error (MQM ≤ -5). P: paragraph-level.

Most evaluations of automatic metrics primarily assess their ability to distinguish between good and bad source-translation pairs (Freitag et al., 2023, 2022b), often overlooking their capacity to discern subtle differences in quality for a given source. Furthermore, in many practical and high-risk applications (*e.g.*, within the medical or legal domains), the main concern is not measuring the *accuracy level* of a translation but determining whether *the translation is accurate and fit for that specific use* (Nida, 1964; Church and Hovy, 1993; Bowker, 2019; Vieira et al., 2021; Mehandru et al., 2023b). While correlations provide valuable insights into the performance of automatic metrics, they do not offer a definitive measure of whether these metrics can reliably confirm translation accuracy.

Hence, in this work, we systematically investigate how existing MT metrics assess high-quality (HQ) correct translations, defined as translations with zero or minor errors only. We find that automatic metrics struggle to distinguish between translations for a given source, especially when comparing HQ translations, with reference-free or quality estimation (QE) metrics achieving close correlation scores to reference-based ones. We further show that current metrics severely overestimate (for non-HQ translations) or underestimate (for HQ translations) translation quality. GEMBA-MQM (Kocmi and Federmann, 2023), a GPT-based QE metric, achieves the highest F1 score in detecting the HQ translations with no errors (HQ-ZERO). However, it also assigns high scores to erroneous GPT-4 translations, suggesting a preferential bias towards the LLM's own outputs. These findings highlight the necessity for more robust evaluation protocols to assess the quality of automatic metrics.

2 How good are current MT systems?

The most reliable way to assess translation quality has been through human evaluations, with several frameworks proposed over the years for this purpose. While earlier works consider two dimensions—adequacy and fluency—with a 5-point Likert scale (King, 1996), subsequent work on direct assessments (DA) considers a single continuous scale of $0 - 100$ (Graham et al., 2017). However, several studies have questioned the credibility of DA-based evaluation (Toral et al., 2018; Läubli et al., 2020; Fischer and Läubli, 2020; Mathur et al., 2020b; Freitag et al., 2021).

Unlike DAs, which assign a numeric score to a translation, the recent Multidimensional Quality Metrics (Burchardt, 2013, MQM) framework relies on explicit error judgments (error types and severity) marked within specific spans of the source-translation pair, providing a more accurate and fine-grained evaluation. Translations receive a score of 0 if they contain no errors, a penalty of -1 for minor errors, and -5 for major errors that impact the usage or understanding of the content.[1]

We present the distribution of gold MQM scores from the WMT23 Metrics task (Freitag et al., 2023), WMT22 Metrics task (Freitag et al., 2022b), and WMT22 Chat Translation task (Farinha et al., 2022) in Table 1. Across settings and language pairs, the percentage of translations achieving a zero MQM score ranges from 19.1% to 63.2%. At least 52.6% translations across language pairs and settings have no major errors (MQM $>$ -5). This shows that a large percentage of the datasets include translations with no or only minor errors, emphasizing the importance of accurately identifying these high-quality translations in the evaluation process.

3 How well do MT metrics assess HQ translations?

We define HQ translations as those that achieve an MQM score > -5, *i.e.*, translations without any major errors according to human evaluators. By definition, these translations do not contain errors that impede their comprehension or usability. We consider a subset of QE and reference-based automatic metrics evaluated by the shared tasks (see App. A for more details).

3.1 How do metrics rank HQ translations?

We first investigate how automatic metrics rank HQ translations, which is particularly relevant today, as these metrics are often used to guide MT training or decoding processes. Recent work employs both reference-based and QE metrics to rerank multiple hypotheses generated by dedicated MT models or large language models (LLMs), aiming to improve translation quality (Fernandes et al., 2022; Freitag et al., 2022a; Farinhas et al., 2023). These metrics are also used to provide quality feedback signals during training, either explicitly in loss signals (Ramos et al., 2023; Yan et al., 2023; He et al., 2024) or implicitly via the creation of preference datasets (Xu et al., 2024b; Yang et al., 2023).

Consider N systems and M source segments. Typically, segment-level correlations are computed between the $N \times M$ translations. However, this differs from the practical setting where metrics are used to rerank several translations for the same source. Therefore, we follow Deutsch et al. (2023) and compute the average correlation between the N

[1] Although MQM includes critical errors—errors that could render a text unusable—they are not marked in many datasets due to their highly contextual interpretation.

	Metric	No-Grouping		Group-by-Src	
		ALL	ALL†	ALL†	HQ
Ref-based	chrF	0.262	0.227	0.267	0.136
	BLEU	0.193	0.190	0.303	0.146
	BERTscore	0.355	0.367	0.325	0.134
	COMET	0.578	0.584	0.461	0.202
	BLEURT-20	0.618	0.603	0.449	0.220
	XCOMET-XL	0.713	0.705	0.461	0.250
	XCOMET-XXL	0.708	0.716	0.481	0.326
	MetricX-23	0.682	0.680	0.450	0.301
	MaTESe	0.591	0.593	0.341	0.254
Ref-free	GEMBA-MQM	0.614	0.621	0.462	0.368
	CometKiwi	0.565	0.561	0.411	0.182
	CometKiwi-XL	0.542	0.550	0.427	0.223
	CometKiwi-XXL	0.525	0.504	0.456	0.327
	MetricX-23-QE	0.683	0.681	0.470	0.292

Table 2: Spearman correlation on WMT23 EN-DE. †: Subsampled to match Group-by-Src HQ's size.

translation scores grouped by the source sentences. We refer to the former setting as **No-Grouping** and the latter as **Group-by-Src**. We also study to what extent these metrics distinguish between HQ translations. As the number of segments with all HQ translations, K, is less than M, we report mean correlations on subsampled datasets (randomly sampled 10 times) that match the sample size, $N \times K$, marked with the symbol † in Table 2. This is motivated by Mathur et al. (2020a), who study how these metrics rank HQ **systems**, where a limited number of samples (typically 4 or 5) was shown to yield unreliable conclusions. However, our focus is on **segment-level** evaluation, where the number of subsampled items is much larger.

Table 2 presents Spearman correlation of automatic metrics with MQM scores for configurations described above on the WMT23 EN-DE dataset (see App. B for other datasets and correlation metrics). We first note that the correlation observed on the entire (No-Grouping ALL) and the subsampled datasets (No-Grouping ALL†) is close, establishing that the observed differences cannot be merely attributed to changes in sample size.

Metrics exhibit only a low-to-fair correlation with human judgments when evaluating translations for the same source. Automatic metrics are less effective in differentiating between good and bad translations for the same source, as evidenced by the drop in correlation from the No-Grouping ALL to the Group-by-Src ALL setting. A possible reason for this disparity lies in how these metrics are typically trained—most metrics are trained to predict translation quality for a given instance (*e.g.*, source-reference-hypothesis trio in Comet or xCOMET). While this can still be useful for ranking two *systems* based on averaged scores across different texts, it may provide limited information for gauging translation quality for different translations of the same source.[2] This highlights the limitations of using automatic metrics as the sole measure of translation quality, particularly in scenarios where fine-grained distinctions between translations of the same source are required.

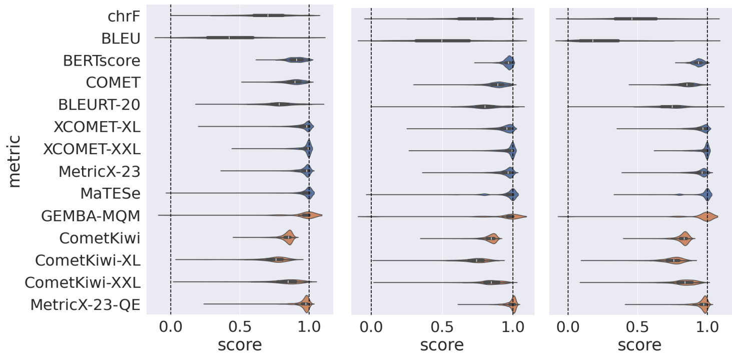
QE metrics are on par with reference-based ones for differentiating translations. QE metrics show promising results in differentiating translations for the same source, often achieving comparable or better correlation than reference-based metrics. For EN-DE, the QE metrics MetricX-23-QE and GEMBA-MQM rank second and third, respectively in the ALL setting, following xCOMET-XXL. When contrasting HQ translations, GEMBA-MQM outperforms all other metrics. The relatively strong performance of QE metrics, particularly in this setting, highlights their potential as valuable tools for translation generation and ranking tasks.

Metrics fail to distinguish HQ translations. There is a consistent drop in correlation scores across all metrics in the HQ relative to the ALL setting, possibly because most translations in the HQ setting receive scores in the narrow range of $(-5, 0]$ and often are tied in quality. Deutsch et al. (2023) show that most metrics struggle to predict translation ties accurately, *i.e.*, give the same score to two translations with similar quality, except for error-predicting metrics like GEMBA-MQM or MaTESe. This decreased correlation from the HQ to the ALL setting has significant implications, especially when they are used to rerank translations produced by strong MT systems. It may result in an artificial boost or bias towards specific systems or outputs, inadvertently prioritizing translations that align well with metric biases but deviate from true quality improvements, as discussed in §3.3.

3.2 How well do metrics detect HQ translations with no errors?

Ranking translations of similar quality is a difficult task, so we also evaluate how automatic metrics score HQ translations with zero MQM scores. (HQ-Zero). We consider normalized scores ≥ 0.99 as

[2]Using contrastive objectives or exposing the metric to multiple translations could potentially help mitigate this issue (Briakou and Carpuat, 2020).



METRIC	EN-DE (1402)			HE-EN (5001)			ZH-EN (11309)		
	P	R	F1	P	R	F1	P	R	F1
xCOMET-XL	**72**	40	51	**78**	17	28	47	28	35
xCOMET-XXL	58	59	58	74	54	62	36	63	46
MaTESe	49	69	58	66	**65**	65	29	**75**	42
MetricX-23	70	33	45	80	16	27	52	11	19
GEMBA-MQM	52	**70**	**60**	71	**65**	**68**	37	77	**50**
MetricX-23-QE	66	14	23	70	64	67	**55**	20	29

Figure 1: **Top:** Metric Scores distribution for HQ-ZERO translations on WMT23. **Bottom:** Precision, recall, and F1.

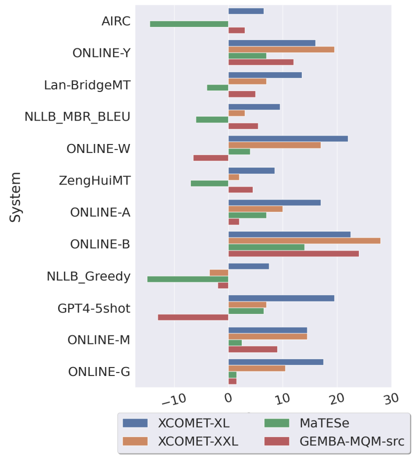


Figure 2: Absolute difference of the number of times a metric assigns a valid score to HQ-ZERO and non HQ-ZERO translations.

valid scores as 1.0 is the highest score a metric should assign to HQ-ZERO translations. Fig. 1 shows the results on WMT23 dataset. See App. C for results in other datasets.

Metric scores have high variance for HQ translations. 9 out of 15 metrics do not assign valid scores to HQ-ZERO translations. Lexical metrics (chrF and BLEU) produce the lowest absolute values, possibly due to over-reliance on a reference translation. Neural metrics trained to regress on DA scores (BLEURT, COMET, and variants) also do not assign valid scores for these translations, likely due to low agreement between DA and MQM scores, as discussed by Freitag et al. (2021).

Metrics over or underestimate translation quality. Metrics that do score these translations within the valid range (xCOMET, MaTESe, MetricX, and GEMBA-MQM), exhibit different tradeoffs between precision (P) and recall (R). For example, while xCOMET-XL and MetricX prioritizes precision, MaTESe and GEMBA-MQM excels at recognizing many HQ-ZERO translations, leading to increased recall. This difference might stem from the specific task each metric is optimized for: while the former predicts sentence-level quality, the latter is optimized to predict word-level error spans. As expected, xCOMET-XXL significantly outperforms xCOMET-XL across all language pairs. Finally, the QE metric, GEMBA-MQM, based on GPT-4, achieves the highest F1 score across all language pairs, demonstrating the capabilities of LLM-based evaluation in more nuanced MT evaluation.

3.3 Which HQ translations are detected?

To study preference bias from metrics towards specific systems, we compute the absolute difference in the number of times a metric assigns a valid score to HQ-ZERO and non-HQ-ZERO translations. Fig. 2 shows that MaTESe equally overestimates translation quality for many systems, as suggested by its high R and low P scores (Fig. 1). GEMBA-MQM frequently assigns zero MQM scores to GPT-4 translations, even when humans identify errors in them. This aligns with concurrent works showing a preference bias of LLMs towards their outputs (Panickssery et al., 2024; Xu et al., 2024c), underscoring the need for a more detailed evaluation to better understand the outputs these metrics prefer and whether they align with human preferences.

4 Conclusions and Future Work

This work systematically investigates how automatic metrics assess HQ translations. We find that current metrics correlate poorly with human judgments when contrasting translations for a given source, with the correlation being even lower for HQ translations. We then study whether metrics can detect HQ translations that attain zero MQM scores (HQ-ZERO) and find that many metrics fail to assign them valid scores. While the GPT-4-based GEMBA-MQM attains the highest F1 for detecting HQ-ZERO, it shows some preference for GPT-4 outputs. Therefore, despite its promise, it is essential to complement GEMBA-MQM with other metrics to ensure robust evaluation.

Limitations

We highlight the main limitations of our work. First, we rely on human MQM annotations as the gold standard for identifying high-quality translations, despite their potential subjectivity and occasional inaccuracy. These annotations are collected for individual translations, and the ratings might vary if annotators were asked to evaluate and compare multiple translations simultaneously.

Second, although our analysis spans multiple datasets across six language pairs (EN-DE, ZH-EN, HE-EN, EN-RU, EN-FR, and EN-PT-BR) and multiple domains, we do not necessarily account for the distribution of high-quality translations across different domains within a dataset. As shown by Zouhar et al. (2024), learned metrics can be sensitive to the domain of evaluation.

Lastly, our analysis in §3.3 identifies one potential bias, but it remains unclear whether automatic metrics have preferential biases towards other output properties such as length, stylistic choices, etc.

References

Sweta Agrawal, Amin Farajian, Patrick Fernandes, Ricardo Rei, and André FT Martins. 2024. Is context helpful for chat translation evaluation? *arXiv preprint arXiv:2403.08314*.

Lynne Bowker. 2019. Fit-for-purpose translation. In *The Routledge handbook of translation and technology*, pages 453–468. Routledge.

Eleftheria Briakou and Marine Carpuat. 2020. Detecting Fine-Grained Cross-Lingual Semantic Divergences without Supervision by Learning to Rank. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1563–1580, Online. Association for Computational Linguistics.

Aljoscha Burchardt. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.

Hannah Béchara, Constantin Orăsan, Carla Parra Escartín, Marcos Zampieri, and William Lowe. 2021. The role of machine translation quality estimation in the post-editing workflow. *Informatics*, 8(3).

Sheila Castilho and Sharon O'Brien. 2017. Acceptability of machine-translated content: A multi-language evaluation by translators and end-users. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 16.

Kenneth W Church and Eduard H Hovy. 1993. Good applications for crummy machine translation. *Machine Translation*, 8:239–258.

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.

Ana C Farinha, M. Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José G. C. de Souza, Helena Moniz, and André F. T. Martins. 2022. Findings of the WMT 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. 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.

Lukas Fischer and Samuel Läubli. 2020. What's the difference between professional human and machine translation? a blind multi-language study on domain-specific MT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 215–224, Lisboa, Portugal. European Association for Machine Translation.

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. 2022a. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

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. 2022b. 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.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

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.

Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. Improving machine translation with human feedback: An exploration of quality estimation as a reward model. *arXiv preprint arXiv:2401.12873*.

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.

Margaret King. 1996. Evaluating natural language processing systems. *Communications of the ACM*, 39(1):73–79.

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

Samuel Läubli, Sheila Castilho, Graham Neubig, Rico Sennrich, Qinlan Shen, and Antonio Toral. 2020. A set of recommendations for assessing human–machine parity in language translation. *Journal of artificial intelligence research*, 67:653–672.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020a. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

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

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023a. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647, Singapore. Association for Computational Linguistics.

Nikita Mehandru, Sweta Agrawal, Yimin Xiao, Ge Gao, Elaine Khoong, Marine Carpuat, and Niloufar Salehi. 2023b. Physician detection of clinical harm in machine translation: Quality estimation aids in reliance and backtranslation identifies critical errors. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11633–11647.

Eugene Albert Nida. 1964. *Toward a science of translating: with special reference to principles and procedures involved in Bible translating*. Brill Archive.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

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

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

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

Miguel Moura Ramos, Patrick Fernandes, António Farinhas, and André FT Martins. 2023. Aligning neural machine translation models: Human feedback in training and inference. *arXiv preprint arXiv:2311.09132*.

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, Daan van Stigt, Marcos Treviso, Luísa Coheur, José GC de Souza, André FT Martins, et al. 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.

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.

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.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. 2021. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society*, 24(11):1515–1532.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. 2024c. Perils of self-feedback: Self-bias amplifies in large language models. *arXiv preprint arXiv:2402.11436*.

Yiming Yan, Tao Wang, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Mingxuan Wang. 2023. BLEURT has universal translations: An analysis of automatic metrics by minimum risk training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5428–5443, Toronto, Canada. Association for Computational Linguistics.

Guangyu Yang, Jinghong Chen, Weizhe Lin, and Bill Byrne. 2023. Direct preference optimization for neural machine translation with minimum bayes risk decoding. *arXiv preprint arXiv:2311.08380*.

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. *arXiv preprint arXiv:2402.18747*.

A Automatic Metrics

We present details about all automatic metrics used across different datasets in Table 3. We refer the reader to the relevant papers (Freitag et al., 2022b, 2023; Agrawal et al., 2024) for more details.

We used the datasets and scores from the WMT 2022 and WMT 2023 Metrics Shared Task campaign, which are available at `https://github.com/google-research/mt-metrics-eval` under the Apache License Version 2.0. For WMT 2022 Chat Shared task human assessments, we used human assessments from `https://github.com/WMT-Chat-task/data-and-baselines/tree/main/data/mqm-annotations` released under a CC-BY-NC license. In our work, we ensured that our usage was consistent with their intended purposes as specified by the licenses.

METRIC	PAPER	INPUT	OUTPUT	TYPE	EVALUATION	DATASET	BASE MODEL
chrF	Popović (2015)	{REF, MT}	$[0\text{-}100] \in \mathbb{R}$	LEXICAL	WMT22, WMT23	-	-
BLEU	Papineni et al. (2002)	{REF, MT}	$[0\text{-}100] \in \mathbb{R}$	LEXICAL	WMT22, WMT23	-	-
BertScore	Zhang et al. (2020)	{REF, MT}	$[0\text{-}1] \in \mathbb{R}$	EMBEDDING	WMT22, WMT23	-	`bert-base-multilingual-cased`
COMET	Rei et al. (2022a)	{SRC, REF, MT}	$[0\text{-}1] \in \mathbb{R}$	LEARNED	WMT23	DA (WMT 2017-2020) + MLQE-PE	`xlm-roberta-large`
BLEURT-20-20	Sellam et al. (2020)	{REF, MT}	$[0\text{-}1] \in \mathbb{R}$	LEARNED	WMT22, WMT23	DA (WMT 2015-2020) + Synthetic	`rembert`
COMET-22∗	Rei et al. (2022a)	{SRC, REF, MT}	$[0\text{-}1] \in \mathbb{R}$	LEARNED	WMT22	DA (WMT 2017-2020) + MLQE-PE + MQM	`xlm-roberta-large`, `infoxlm-large`
MetricX-22	-	{REF, MT}	$[-25,0], \mathbb{R}$	LEARNED	WMT22	-	`30B mT5`
MetricX-23	Juraska et al. (2023)	{REF, MT}	$[-25,0] \in \mathbb{R}$	LEARNED	WMT23	DA (WMT 2015-2020) + MQM (WMT 2020-2021) + Synthetic	`mT5-XXL`
xCOMET∗	Guerreiro et al. (2023)	{SRC, REF, MT}	$[0\text{-}1] \in \mathbb{R}$	LEARNED	WMT23	DA (WMT 2017-2020) + MLQE-PE + MQM (WMT 2020-2021; IndicMT, DEMETR) + Synthetic	`XLM-RoBERTa-XL`, `XLM-RoBERTa-XXL`
MaTESe	Perrella et al. (2022)	{REF, MT}	$[-25,0] \in \mathbb{Z}$	textscLearned	WMT22	MQM (WMT 2020-2021)	`XLM-RoBERTa`, `BART`
MaTESe	-	{REF, MT}	$[-25,0] \in \mathbb{Z}$	textscLearned	WMT23	MQM (WMT 2020-2022)	`DeBERTa`, `InfoXLM`
GEMBA-MQM	Kocmi and Federmann (2023)	{SRC, MT}	$[-25,0] \in \mathbb{Z}$	LLM-based	WMT23	-	`GPT4`
CometKiwi-22∗	Rei et al. (2022b)	{SRC, MT}	$[0\text{-}1] \in \mathbb{R}$	LEARNED	WMT22	DA (WMT 2017-2020) + MLQE-PE + MQM †	`rembert`, `infoxlm-large`
CometKiwi-23∗	Rei et al. (2023)	{SRC, MT}	$[0\text{-}1] \in \mathbb{R}$	LEARNED	WMT23	DA (WMT 2017-2020) + MLQE-PE + MQM †	`rembert`, `infoxlm-large`
MetricX-23-QE	Juraska et al. (2023)	{SRC, MT}	$[-25,0] \in \mathbb{R}$	LEARNED	WMT23	DA (WMT 2015-2020) + MQM (WMT 2020-2021) + Synthetic	`mT5-XXL`
MaTESe-QE	Perrella et al. (2022)	{SRC, MT}	$[-25,0] \in \mathbb{Z}$	textscLearned	WMT22	MQM (WMT 2020-2021)	`XLM-RoBERTa`, `BART`
xCOMET-QE∗	Guerreiro et al. (2023)	{SRC, MT}	$[0\text{-}1] \in \mathbb{R}$	LEARNED	WMT23	DA (WMT 2017-2020) + MLQE-PE + MQM (WMT 2020-2021; IndicMT, DEMETR) + Synthetic	`XLM-RoBERTa-XL`, `XLM-RoBERTa-XXL`

Table 3: Details about the automatic metrics considered in our paper. $*$: submission is an ensemble; †: {SRC, REF} pairs are also added to the training data.

B Ranking results

Tables 4 and 5 report the Spearman and Pearson correlation results for WMT23 EN-DE, respectively. Tables 6 and 7 show the Spearman Correlation for the WMT22 and WMT23 datasets, respectively. We do not perform this analysis on chat data because the number of systems is ≤ 5.

METRIC	NO-GROUPING			NO-GROUPING †			GROUP-BY-SRC		
	ALL	HQ	Δ	ALL	HQ	Δ	ALL†	HQ	Δ
chrF	0.262	0.137	−0.124	0.227 ±0.030	0.132 ±0.022	−0.094	0.267 ±0.050	0.136	−0.131
BLEU	0.193	0.094	−0.099	0.190 ±0.032	0.087 ±0.022	−0.103	0.303 ±0.056	0.146	−0.156
BERTscore	0.355	0.190	−0.165	0.367 ±0.039	0.183 ±0.032	−0.184	0.325 ±0.035	0.134	−0.191
COMET	0.578	0.385	−0.194	0.584 ±0.024	0.390 ±0.031	−0.194	0.461 ±0.041	0.202	−0.259
BLEURT-20	0.618	0.357	−0.262	0.603 ±0.020	0.357 ±0.033	−0.246	0.449 ±0.043	0.220	−0.229
XCOMET-XL	0.713	0.454	−0.259	0.705 ±0.020	0.449 ±0.018	−0.256	0.461 ±0.030	0.250	−0.211
XCOMET-XXL	0.708	0.399	−0.309	0.716 ±0.020	0.382 ±0.032	−0.335	0.481 ±0.041	0.326	−0.155
MetricX-23	0.682	0.433	−0.249	0.680 ±0.018	0.446 ±0.027	−0.233	0.450 ±0.043	0.301	−0.149
MaTESe	0.591	0.353	−0.238	0.593 ±0.028	0.370 ±0.044	−0.223	0.341 ±0.042	0.254	−0.087
quality estimation									
GEMBA-MQM	0.614	0.345	−0.269	0.621 ±0.027	0.358 ±0.028	−0.263	0.462 ±0.044	0.368	−0.094
CometKiwi	0.565	0.286	−0.279	0.561 ±0.019	0.268 ±0.021	−0.293	0.411 ±0.044	0.182	−0.229
CometKiwi-XL	0.542	0.240	−0.302	0.550 ±0.023	0.254 ±0.032	−0.296	0.427 ±0.029	0.223	−0.204
CometKiwi-XXL	0.525	0.236	−0.289	0.504 ±0.031	0.244 ±0.032	−0.260	0.456 ±0.029	0.327	−0.129
MetricX-23-QE	0.683	0.425	−0.258	0.681 ±0.012	0.439 ±0.027	−0.242	0.470 ±0.028	0.292	−0.177

Table 4: Spearman correlation on WMT23 EN-DE. †: Subsampled to match GROUP-BY-SRC HQ's sample size.

METRIC	NO-GROUPING			NO-GROUPING †			GROUP-BY-SRC		
	ALL	HQ	Δ	ALL	HQ	Δ	ALL†	HQ	Δ
chrF	0.232	0.112	−0.120	0.244 ±0.028	0.121 ±0.028	−0.123	0.322 ±0.041	0.124	−0.198
BLEU	0.192	0.086	−0.106	0.210 ±0.029	0.079 ±0.025	−0.131	0.297 ±0.049	0.148	−0.149
BERTscore	0.325	0.150	−0.175	0.331 ±0.038	0.148 ±0.031	−0.182	0.363 ±0.043	0.150	−0.213
COMET	0.432	0.337	−0.095	0.421 ±0.037	0.367 ±0.031	−0.055	0.513 ±0.044	0.266	−0.246
BLEURT-20	0.484	0.324	−0.160	0.488 ±0.021	0.308 ±0.024	−0.180	0.469 ±0.047	0.245	−0.223
XCOMET-XL	0.680	0.414	−0.266	0.680 ±0.028	0.409 ±0.040	−0.272	0.510 ±0.054	0.359	−0.150
XCOMET-XXL	0.695	0.362	−0.333	0.688 ±0.019	0.355 ±0.038	−0.333	0.484 ±0.068	0.385	−0.098
MetricX-23	0.585	0.406	−0.179	0.576 ±0.023	0.406 ±0.025	−0.169	0.512 ±0.024	0.371	−0.141
MaTESe	0.554	0.238	−0.316	0.547 ±0.035	0.221 ±0.032	−0.325	0.345 ±0.045	0.253	−0.092
quality estimation									
GEMBA-MQM	0.502	0.223	−0.279	0.497 ±0.027	0.238 ±0.021	−0.260	0.485 ±0.055	0.386	−0.099
CometKiwi	0.475	0.210	−0.265	0.476 ±0.037	0.198 ±0.049	−0.277	0.458 ±0.057	0.226	−0.232
CometKiwi-XL	0.446	0.185	−0.262	0.445 ±0.033	0.198 ±0.032	−0.247	0.499 ±0.041	0.328	−0.171
CometKiwi-XXL	0.417	0.171	−0.245	0.411 ±0.024	0.167 ±0.040	−0.244	0.531 ±0.040	0.378	−0.152
MetricX-23-QE	0.626	0.371	−0.255	0.640 ±0.036	0.372 ±0.029	−0.268	0.536 ±0.048	0.407	−0.129

Table 5: Pearson correlation on WMT23 EN-DE. †: Subsampled to match GROUP-BY-SRC HQ's sample size.

	WMT23 HE-EN				WMT23 ZH-EN			
	NO-GROUPING †		GROUP-BY-SRC		NO-GROUPING †		GROUP-BY-SRC	
METRIC	All	HQ	All†	HQ	All	HQ	All†	HQ
chrF	0.299	0.140	0.298	0.144	0.067	0.012	0.220	0.162
BLEU	0.248	0.145	0.270	0.161	0.129	0.065	0.190	0.139
BERTscore	0.391	0.210	0.368	0.191	0.269	0.129	0.273	0.154
COMET	0.485	0.226	0.383	0.167	0.457	0.268	0.315	0.183
BLEURT-20	0.459	0.216	0.379	0.173	0.434	0.241	0.332	0.189
XCOMET-XL	0.511	0.255	0.362	0.147	0.608	0.405	0.334	0.185
XCOMET-XXL	0.528	0.260	0.381	0.140	0.607	0.364	0.373	0.219
MetricX-23	0.549	0.258	0.357	0.171	0.603	0.408	0.339	0.202
MaTESe	0.415	0.207	0.353	0.266	0.467	0.277	0.322	0.216
quality estimation								
GEMBA-MQM	0.493	0.245	0.420	0.227	0.580	0.358	0.423	0.264
CometKiwi	0.459	0.225	0.309	0.106	0.533	0.328	0.333	0.160
CometKiwi-XL	0.434	0.184	0.348	0.181	0.532	0.302	0.334	0.170
CometKiwi-XXL	0.468	0.213	0.389	0.202	0.504	0.288	0.352	0.161
MetricX-23-QE	0.495	0.235	0.307	0.126	0.621	0.411	0.322	0.159
XCOMET-QE-Ensemble	0.504	0.233	0.345	0.160	0.631	0.377	0.347	0.177

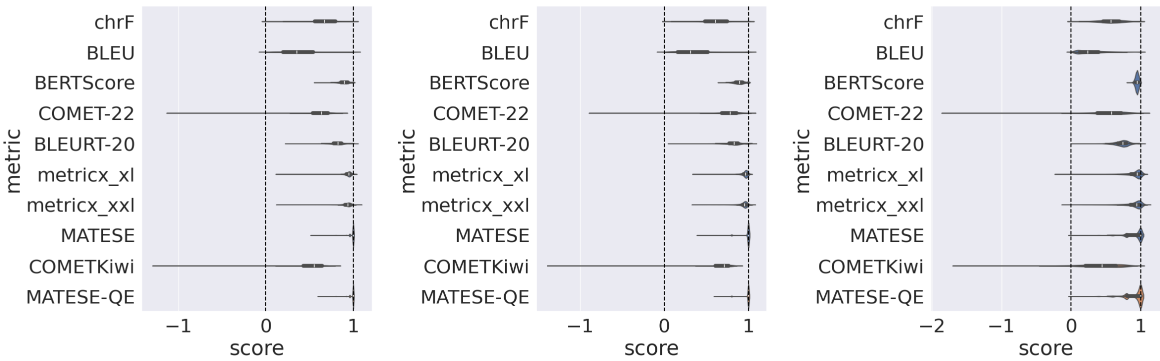
Table 6: Spearman correlation on WMT23 (HE-EN and ZH-EN). †: Subsampled to match GROUP-BY-SRC HQ's sample size.

	WMT22 EN-DE				WMT22 EN-RU				WMT22 ZH-EN			
	NO-GROUPING †		GROUP-BY-SRC		NO-GROUPING †		GROUP-BY-SRC		NO-GROUPING †		GROUP-BY-SRC	
METRIC	All	HQ	All†	HQ	All	HQ	All†	HQ	All	HQ	All†	HQ
chrF	0.296	0.214	0.242	0.206	0.235	0.161	0.237	0.161	0.199	0.069	0.189	0.096
BLEU	0.233	0.176	0.221	0.210	0.194	0.161	0.198	0.127	0.200	0.086	0.146	0.089
BERTScore	0.318	0.244	0.239	0.207	0.265	0.210	0.240	0.158	0.428	0.189	0.265	0.155
COMET-22	0.497	0.392	0.358	0.314	0.534	0.387	0.394	0.282	0.428	0.189	0.265	0.155
BLEURT-20	0.467	0.346	0.352	0.283	0.483	0.342	0.354	0.257	0.488	0.194	0.305	0.170
MetricX-XL	0.499	0.379	0.395	0.349	0.511	0.392	0.379	0.290	0.550	0.253	0.314	0.210
MetricX-XXL	0.490	0.377	0.370	0.304	0.561	0.430	0.402	0.338	0.554	0.260	0.303	0.204
MaTESe	0.387	0.296	0.356	0.349	0.315	0.236	0.321	0.281	0.477	0.243	0.251	0.222
quality estimation												
CometKiwi	0.404	0.300	0.273	0.223	0.482	0.341	0.306	0.228	0.488	0.223	0.263	0.205
MaTESe-QE	0.294	0.236	0.314	0.316	0.258	0.184	0.268	0.256	0.412	0.214	0.212	0.208

Table 7: Spearman correlation on WMT22 (EN-DE, EN-RU, annd ZH-EN). †: Subsampled to match GROUP-BY-SRC HQ's sample size.

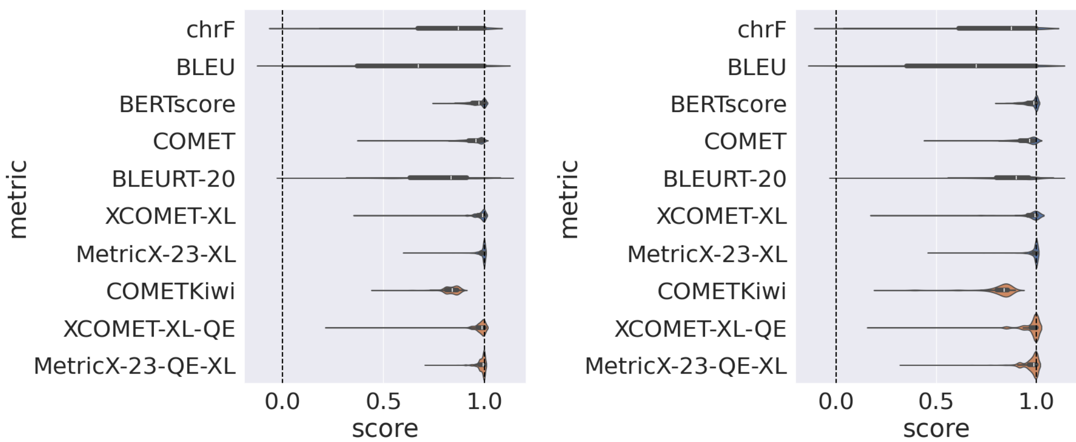
C HQ-ZERO Detection Results

We present the results for the detection task on the WMT22 Metrics and Chat datasets in Figures 3 and 4, respectively.



METRIC	EN-DE			EN-RU			ZH-EN		
	P	R	F1	P	R	F1	P	R	F1
MaTESe	61	86	71	48	94	63	68	53	60
MaTESe-QE	58	87	70	46	95	62	64	55	59

Figure 3: **Top:** Scores distribution for HQ-ZERO translations on WMT22. **Bottom:** Precision, recall, and F1.



METRIC	EN-XX			XX-EN		
	P	R	F1	P	R	F1
chrF	88	38	53	92	42	58
BLEU	88	38	53	93	42	58
BERTScore	93	23	37	94	27	42
XCOMET-XL	75	33	46	87	38	53
MetricX-23-XL	76	64	69	87	62	72
XCOMET-XL-QE	66	29	40	84	49	62
MetricX-23-QE-XL	76	45	56	80	35	49

Figure 4: **Top:** Scores distribution for HQ-ZERO translations on WMT22 Chat. **Bottom:** Precision, recall, and F1.