MSLC25: Metric Performance on Low-Quality Machine Translation, Empty Strings, and Language Variants

Rebecca Knowles **Samuel Larkin** **Chi-kiu Lo** 羅致翹
Digital Technologies Research Centre
National Research Council Canada (NRC-CNRC)
{rebecca.knowles,samuel.larkin,chikiu.lo}@nrc-cnrc.gc.ca

Abstract

In this challenge set, we examine how automatic metrics for machine translation perform on a wide variety of machine translation output, covering a wider range of quality than the WMT submissions. We also explore metric results on specific types of corner cases, such as empty strings, wrong- or mixed-language text, and more. We primarily focus on Japanese–Chinese data, with some work on English and Czech.

1 Introduction

This paper describes a challenge set submitted to the Challenge Set Subtask of the (Unified) Machine Translation Evaluation shared task at the 2025 Conference on Machine Translation (WMT); we focus primarily on segment-level quality score prediction with a brief preliminary note on word-level error detection and span annotation. For this third iteration of the Metric Score Landscape Challenge (MSLC),[1] we run a smaller set of experiments. We once again focus on Japanese→Chinese news translation for the task of exploring the range of low- to mid-quality MT (as compared to the high-quality MT systems submitted to WMT). We also include analysis of empty strings (as in past iterations), mixed- and wrong-language text, and a preliminary note on English spelling variation; for these tasks, we use a mix of Japanese→Chinese data, English language data, and Czech→English data. Our approach, particularly for low- to mid-quality MT and for empty strings, demonstrates a low-cost way to test metrics on the wider quality landscape. We encourage developers of metrics to run such evaluations themselves prior to releasing metrics. For developers of metrics who are unable to run such evaluations themselves, we call on them to explicitly declare that their released metrics have not been tested on low- or mid-quality MT and should not be used for such cases without additional testing.

[1] MSLC data and additional figures can be found at https://github.com/nrc-cnrc/MSLC.

2 Prior and Related Work

This work describes the third MSLC challenge set. In Lo et al. (2023b), the first iteration, our intent was to focus primarily on low- and mid-quality MT output across four language pairs. The process of evaluation that year brought to light two other issues: the scores that metrics assigned to empty strings (as some MT system submission that year included empty strings, providing a natural experimental set) and "universal scores" (scores that were assigned very frequently by certain metrics) similar to the "universal translations" described in Yan et al. (2023). The second iteration of MSLC, Knowles et al. (2024), continued the work on low- and mid-quality MT, with the addition of experiments on empty strings, mixed- and wrong-language text, and language variants (in that instance, Spanish language terminology that differs across language variants). In concurrent work to MSLC24, Zouhar et al. (2024) proposed COMET-specific mitigations to many of the issues observed in both papers: incorporating language ID in order to mitigate issues with mixed- or wrong-language text, using signatures in the spirit of `sacreBLEU` (Post, 2018) to help explain variations in metric output, and the issue of empty strings.

There is a tradition of challenge sets targeting specific linguistic phenomena for MT (Isabelle et al., 2017; Burlot and Yvon, 2017; Guillou et al., 2018; Rios et al., 2018; Stanovsky et al., 2019, i.a.), while challenge sets for evaluation (i.e., challenge sets targeted at metrics) tend to be relatively newer (see, i.a., the descriptions of the challenge sets at the Metrics shared tasks: Freitag et al., 2022, 2023, 2024). Of these challenge sets, Amrhein et al. (2022, 2023) also explore wrong-language text (among 68 phenomena across a large number of language pairs in the ACES challenge set),

Proceedings of the Tenth Workshop on Machine Translation (WMT), pages 945–956
November 8–9, 2025.

noting particular issues for reference-free metrics, similar to our observations.

Our work is situated more broadly in the area of MT evaluation and corner cases for MT evaluation. While there is prior work focusing on specific metrics and corner cases (Hanna and Bojar, 2021; Amrhein and Sennrich, 2022; Yan et al., 2023; Zouhar et al., 2024, i.a.), submitting this challenge set to the shared task permits us to examine and compare performance on corner cases and MT quality ranges across metrics in a controlled environment. Importantly, readers should note that this paper is a description of a challenge set submitted to a shared task, rather than a complete, in-depth exploration of all the areas on which it touches. We note, in both the limitations section and throughout the work, that there are some components that represent established evaluations (MSLC-A and the empty strings work) while other components are presented as initial proof-of-concept experiments to determine whether future in-depth evaluation is indicated (these smaller preliminary experiments should not be used to draw sweeping conclusions).

3 Data

The MSLC challenge set is divided into two main components: MSLC-A (which covers low-quality and mid-quality MT) and MSLC-B (which targets specific corner cases and potential challenges for metrics).

3.1 MSLC-A

The MSLC-A portion of the challenge set focuses on covering a range of MT quality, from extremely low quality (incomprehensible output) to mid-quality output. The intention is to fill some of the gaps in evaluation of new metrics, which are typically tested at WMT on high-quality systems only, despite the fact that they may go on to be used for a wider range of quality in practice.

We use Japanese→Chinese systems from Knowles et al. (2024). The MT models were all constrained (as per the 2024 WMT General Task rules) NMT models built using Sockeye version `3.1.31` (Hieber et al., 2022) and trained on WMT training data; for a more detailed description of the systems, see Larkin et al. (2024). We translate and use only the News portion of the 2025 WMT General Task data.

There are six MT systems in this part of the challenge set. The lowest-quality system is indicated with the letter A, and the quality approximately increases as the system labels proceed alphabetically. Using the same process described in Larkin et al. (2024) and Knowles et al. (2024), these low- to mid-quality outputs were produced by translating the same source text[2] using early checkpoints saved during model training, with the lowest quality produced by the early checkpoints (i.e., when the system produced nonsensical and repetitive output) and the mid-quality outputs produced by later checkpoints. These low- to mid-quality MT system outputs were ranked by *BLEU* and manually examined (on a subset of the data) by an author fluent in the target language to confirm their increasing quality.[3] We run a limited version of the MSLC-A experiments in this edition of the challenge set, without submission of the systems to the General Task at WMT (i.e., we do not have human evaluations that indicate the magnitude of the gap, if any, between our highest-performing mid-quality MT system and the lowest-performing submitted system).

3.2 MSLC-B

For the MSLC-B portion of the challenge set, we focus on three different types of edge cases for metrics: empty strings, mixed- and wrong-language text, and English spelling variants.

In past iterations of MSLC, we observed that differences in what a metric treats as a "document" can have an impact on the scores it assigns in our test sets. In an effort to ensure that document-level metrics did not mix together the contrastive examples we were having scored, we appended strings to the document IDs to identify these as "separate documents" where appropriate (i.e., when two contrastive examples were from the same document, they might be assigned as "[DOCID]-1" and "[DOCID]-2" so that the metric should treat them as separate documents).

3.2.1 Empty Strings

Given past observations (Lo et al., 2023b; Knowles et al., 2024) of unusual outputs related to empty

[2]Text was translated at the segment or sentence level and then re-collected into documents.

[3]While it may be desirable to perform more formal evaluation, the lowest-quality systems are of such low quality as to be visibly "nonsense" even to non-speakers of the language as well, with a clear trend of improvement. Thus, while we do not have MQM or other manual scores to rank these, we can be confident in the overall trend, and particularly confident that, e.g., system A is a substantially worse system than system D.

strings, such as surprisingly high scores, we once again examine this topic. We use a small Japanese→Chinese dataset for this: 10 punctuation characters, 10 words, 10 phrases, and 10 segments (sentences or larger); all except for the punctuation are selected from the news domain portion of the WMT 2025 General Task test data. For each of these, we explore the case where we have an empty source and reference and a non-empty hypothesis (representing overgeneration: an MT system producing something from nothing) and the case where there is an empty hypothesis with a full source and reference (undergeneration: an MT system producing nothing when it should have produced something).

3.2.2 Mixed- and Wrong-Language Text

For metrics that rely on multilingual embeddings and ones that do not take into account the intended source and target language, there is a risk of returning high scores for wrong-language output. As in Knowles et al. (2024), we examine metric scores when presented with wrong-language or mixed-language hypotheses. In this case, we take advantage of the multi-way parallel test sets (pivoted on English) by using segments from English news test data with its Chinese and Japanese translations. We run these experiments on 18 segments with Japanese→Chinese as the intended language pair and translation direction, using the Japanese data as the source, the Chinese data as the reference, the English data as the wrong-language hypothesis, and a pseudo-codeswitched mix of English and Chinese as the mixed-language hypothesis.[4]

3.2.3 English Language Spelling Variants

While recent iterations of WMT have included more regional specifications regarding language variants, variations in English have been overlooked. We use a word list of common UK (en_GB) and US English (en_US) spelling differences[5] to select segments from the Czech→English news test data that contain words with the potential for different spellings.[6] We then automatically produce two versions of each of these 20 English segments, one with standard UK spellings and one with standard US spellings, and confirm them with manual examination. The English sentences are otherwise identical: minimal pairs where the only difference is the spelling of the words of interest. We then inverted the translation direction, treating Czech as the source and English as the target. The official submission format for the Metrics Challenge Set subtask included a field for language ID for the target; we submitted versions that included the region (en_US or en_UK) and one that did not (en). This should enable us to see whether some metrics utilize this information and, for those that do not, whether there is a bias towards a particular language variant. Of note, this is most relevant to reference-free metrics, as the reference is either an exact match to the hypothesis or identical except for the spelling of the words of interest. This is a small preliminary experiment to determine whether future large-scale evaluation of this topic may be fruitful.

4 Metrics

We focus on analyzing the scores produced by the baseline metrics and the primary submissions. There are 9 baseline metrics (including 2 "sentinel" metrics designed to scrutinize the metric meta-evaluation process) and 5 primary metrics that participated in portions of our challenge set for segment-level evaluation. One baseline and 2 primary metrics participated in the error span detection portion of our challenge set.

The segment-level score baselines are *BLEU* (Papineni et al., 2002), *spBLEU* (NLLB Team et al., 2022), *chrF* (Popović, 2015), *BERTScore* (Zhang et al., 2020), *COMET-22* (Rei et al., 2022a), *CometKiwi* (Rei et al., 2022b), *YiSi-1* (Lo, 2019), *sentinel-cand* and *sentinel-src* (Perrella et al., 2024).

For the segment-level quality score prediction task, five metrics participated in our experiments. *MetricX-25* (Juraska et al., 2025), an updated version of *MetricX*, is an encoder-only regression model initialized from Gemma 3 (Team et al., 2025) 12B and fine-tuned on publicly available DA and MQM scores from WMT 2015–23. *mr7.2.1* (Hrabal et al., 2025) is based on the Gemma 3 27B IT model and is prompted with the DSPy (Khattab et al., 2024) framework and its MIPROv2 opti-

[4]This mixed-language data was produced manually by an author fluent in the languages and is intended to contain the full semantic content of the text in such a way that it could be read by someone who speaks both languages and be perceived as similar to naturally generated by code-switching speakers.

[5]https://github.com/hyperreality/American-British-English-Translator/blob/master/data/american_spellings.json

[6]In some cases, we shortened the segments to more tightly focus on sentences containing the words of interest (shortening both the English and Czech sides of the sentence pair).

mizer. *Polycand-2* (Züfle et al., 2025) is a COMET-based metric that incorporates two alternative translations of the same source segment (provided by other translation systems) to better contextualize and assess the quality of the translation being scored. *rankedCOMET* (Maharjan and Shrestha, 2025) is a COMET-based metric post-processed with rank normalization for each language pair. *UvA-MT* (Wu and Monz, 2025) calibrates quality estimation and likelihood on the Gemma 3 12B IT model, then directly uses the token average likelihood as a metric for quality estimation.

For the error span detection task, three systems participated in our MSLC-B experiments: baseline *XCOMET* (Guerreiro et al., 2024) and the two primary submissions of *AIP1* (Yeom et al., 2025) and *GemSpanEval* (Juraska et al., 2025). *AIP1* uses the OpenAI o3 (OpenAI, 2025) reasoning model and its structured-output mode to detect translation errors at the span level. *GemSpanEval* is based on the Gemma 3 27B model and is finetuned to predict MQM error spans.

Some metrics use reference translations in the process of producing their scores, while others do not. Throughout the remainder of the paper we indicate the 3 baseline and 4 primary reference-free (QE) metrics—those that do not use the reference translation in their scoring—with an asterisk before the metric name. There are 3 reference-free primary metrics (**mr7.2.1*, **Polycand-2*, **UvA-MT*) in the segment-level score prediction task and 1 (**AIP1*) in the error span prediction task.

5 MSLC-A Results and Plots

Interpretation note of caution: our submitted MSLC-A challenge set was scored at the document level by the automatic metrics, while the submitted primary MT systems from the general task were scored at a sub-document segment level by the automatic metrics. In order to be able to compare these, we have averaged the segment-level scores to produce a document-level score for each document. We note that this may not always be identical to the score that the metric would have assigned had it scored the full document directly, and caution should therefore be taken when drawing conclusions.

Figure 1 shows system average scores for the MSLC-A systems (cool colours, left) and the systems submitted the the WMT General MT task, computed only over the News data. These averages are computed from the document-level scores, which in the case of MSLC-A data were produced directly by the metrics and which in the case of the submitted systems were produced as an average of segment-level scores. We find several points of interest. Three of the metrics, **COMETKiwi22*, *MetricX-25*, and **UvA-MT* have some difficulty with the rankings of the lowest-quality MSLC-A systems. In particular, **UvA-MT* ranks the worst system (whose output is almost entirely nonsensical and unreadable) as similar in quality to the high-quality WMT submissions. This indicates that these metrics—for this language pair at least—may not be trustworthy metrics to use when trying to evaluate low-quality MT. Metric users should consider alternative choices of metrics if they have reason to believe that their MT output may be of low quality or mid-range quality. This could also have an impact on MT systems trained using these metrics, particularly in the early stages of training.

We also observe that some metrics (**Polycand-2*, *COMET22*, i.a.) devote a very small portion of their metric's space of possible scores to the high-quality systems, with a wider range of scores for the low-quality systems, while others like *chrF* distribute the score range more evenly. This may impact how useful a metric is for distinguishing between systems of different levels of quality, and may also play a role in human interpretation of metric score differences; for broader discussion of metric score differences and human interpretations thereof, see Mathur et al. (2020) and Lo et al. (2023a), i.a. There is not an inherent right answer to how a metric should use its score space, rather, it is tied to the intended use of the metric.

We observe that some metrics show overlap or near overlap between the best of the MSLC systems and the lowest-scoring of the submitted systems while others show a large gap between them. Since we do not have full human annotations available, we cannot make any claims about whether there would be an overlap or a gap based on human evaluation.

Figure 2 provides another way to visualize the metric scores, showing histograms of scores assigned to each system along the diagonal, and scatterplots showing correlations between metrics on the off-diagonals. The metrics that give higher-than-expected scores to the low-quality systems once again stand out where they do not correlate with systems that rank the low-quality systems as

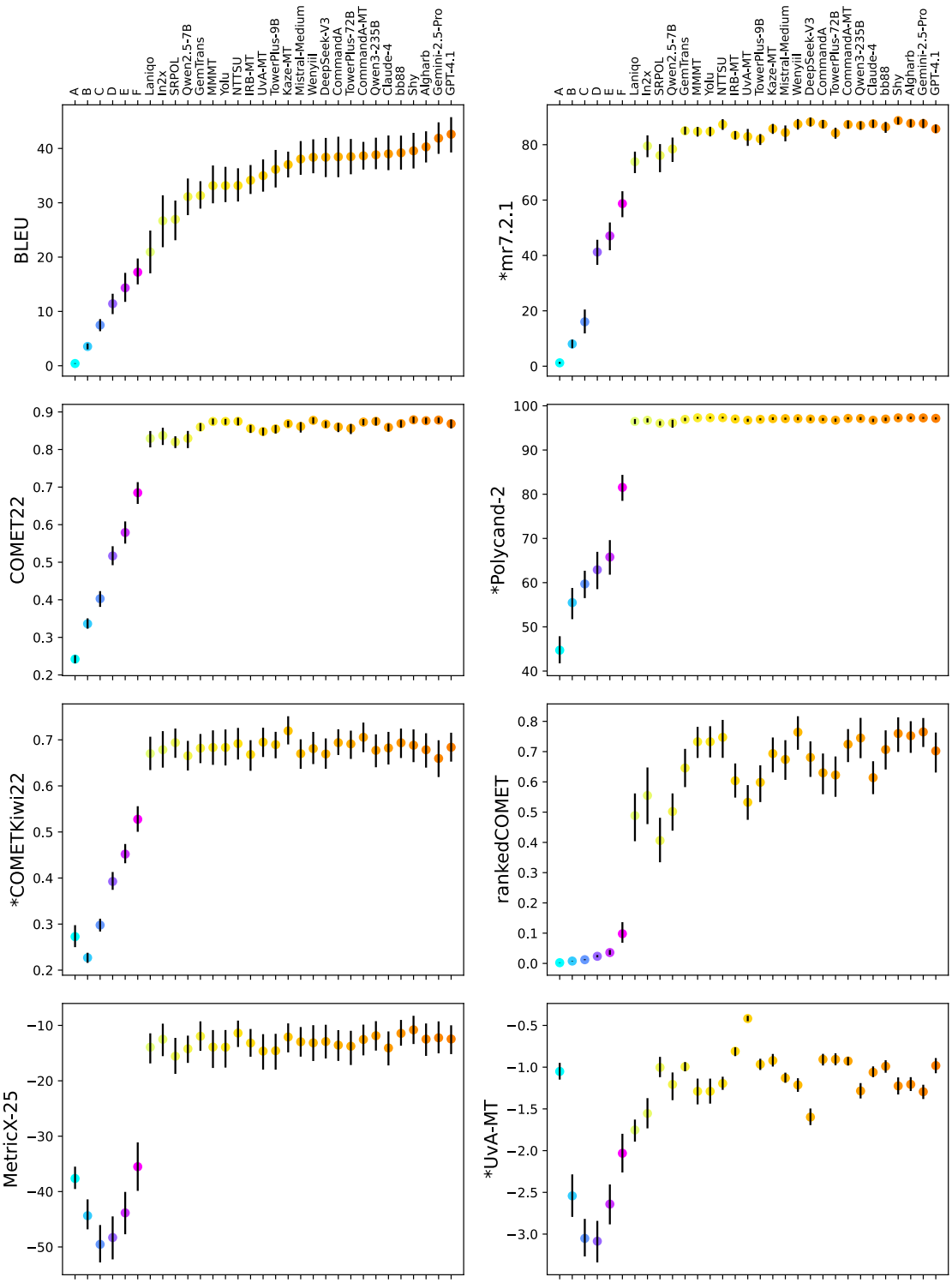

Figure 1: System average scores for Japanese→Chinese. MSLC systems (cool colours, left) are ordered by BLEU score and brief manual examination; WMT submitted systems are ranked by average BLEU score.

expected. We can also use the histograms to gain a better understanding of the score distributions assigned to the data, such as the very low scores and comparatively small score range assigned by *rankedCOMET* to the low-quality systems, as compared to other metrics, or the somewhat bimodal score distribution from **mr7.2.1*. In past iterations of MSLC, this type of figure was particularly useful in highlighting unusual properties of some of the metrics, such as discretizing the score space or

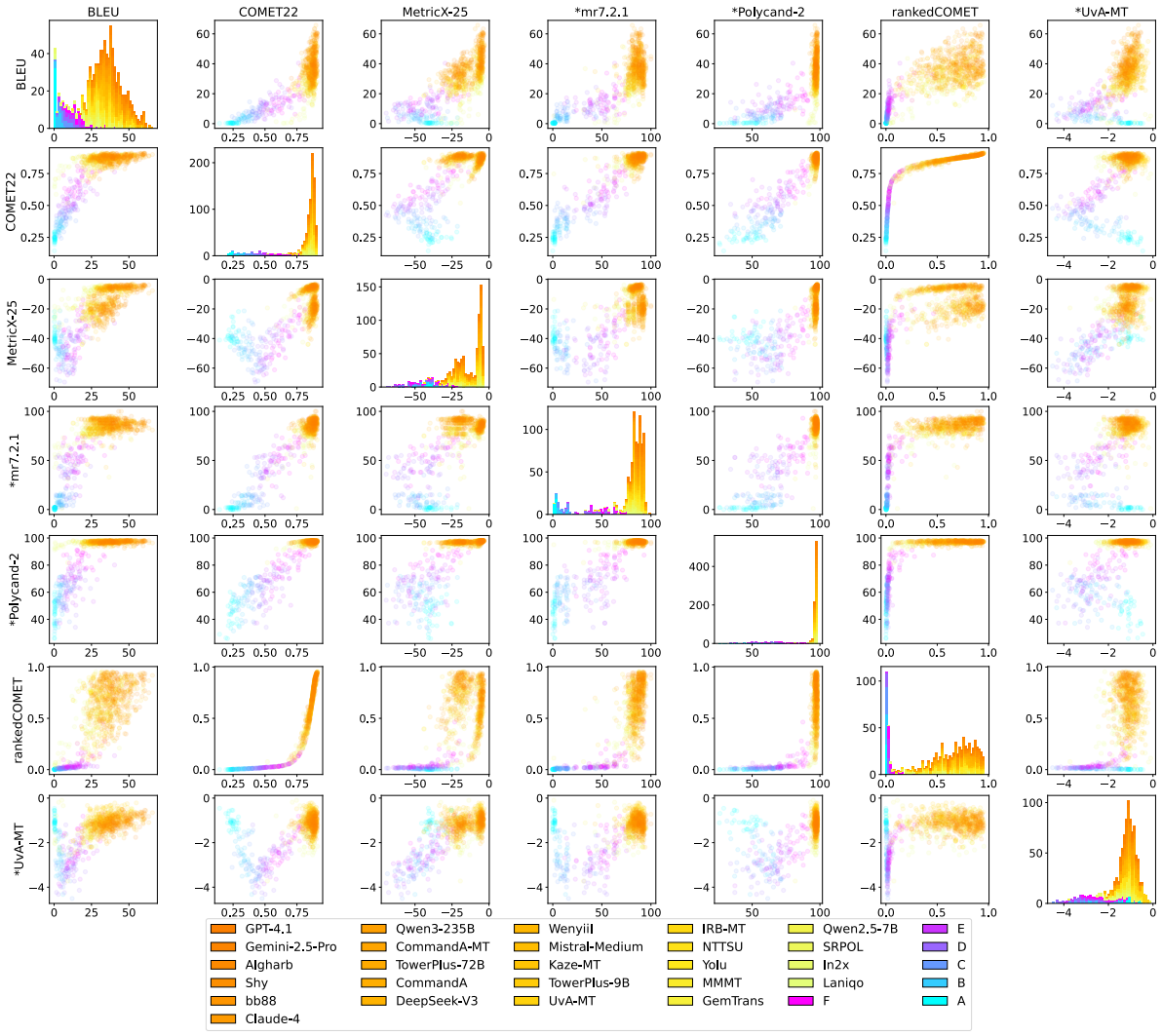

Figure 2: Matrix of segment-level scores for Japanese→Chinese. Along the diagonal are stacked histograms of segment scores across the challenge set (cool colours/bottom) and submitted WMT systems (warm colours/top). The off-diagonal entries are scatterplots where each point is a single document positioned according to the score assigned to it by row and column metrics; each point is coloured according to the same colours as the histogram.

assigning specific scores very frequently ("universal scores"). We do not observe such results in this year's set of metrics.

Due to scheduling constraints, we did not receive the human annotation scores in time to display those in these figures; we plan to incorporate them into final additional figures on the MSLC website (https://github.com/nrc-cnrc/MSLC).

6 MSLC-B Results and Plots

6.1 Empty Strings

In Figure 3 we show the scores assigned by the five primary submission metrics to punctuation, words, phrases, and segments (sentences to documents) when those are paired with an empty source and reference. The vertical red lines indicate the minimum and maximum scores assigned by the same metric to all WMT General Task primary submissions on the News portion of the data; since different metrics use different score ranges, this is used to provide the reader with some context about where the scores for these corner cases fall in comparison to scores assigned to more usual MT output. These empty string examples are fairly extreme examples of MT failures; string-based metrics like *BLEU* would assign them scores of 0.

We see several types of responses. The metric **mr7.2.1* assigns its lowest score to all of these, similar to what we observe from metrics like *chrF* and *BLEU* (not shown in figure); this is arguably what we would expect, since an empty source should produce an empty hypothesis. The results from

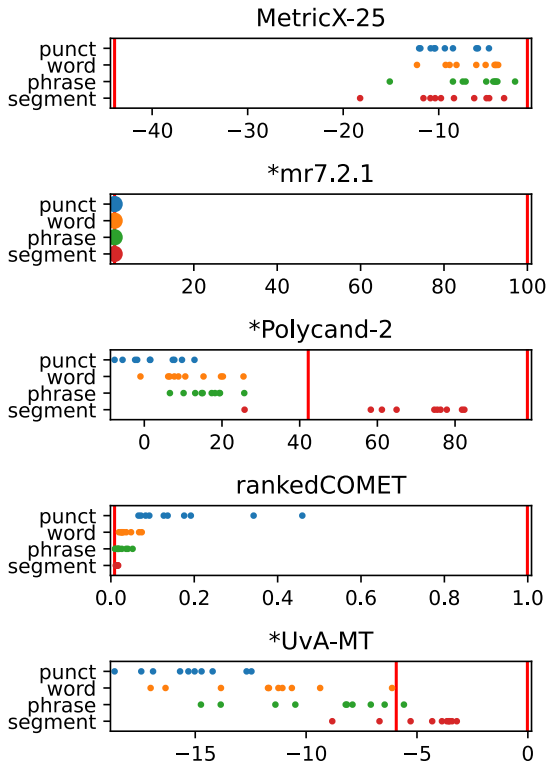

Figure 3: Japanese→Chinese scores assigned to text when paired with empty source and reference. Where multiple strings receive the same score, this is indicated by proportionally increased dot size (in the case of **mr7.2.1*, all strings received the same score of 0, as indicated by the large dots on top of the left red vertical line). Red vertical lines indicate the minimum and maximum scores assigned over all Japanese→Chinese WMT News primary submission data. Asterisks indicate reference-free (QE) metrics.

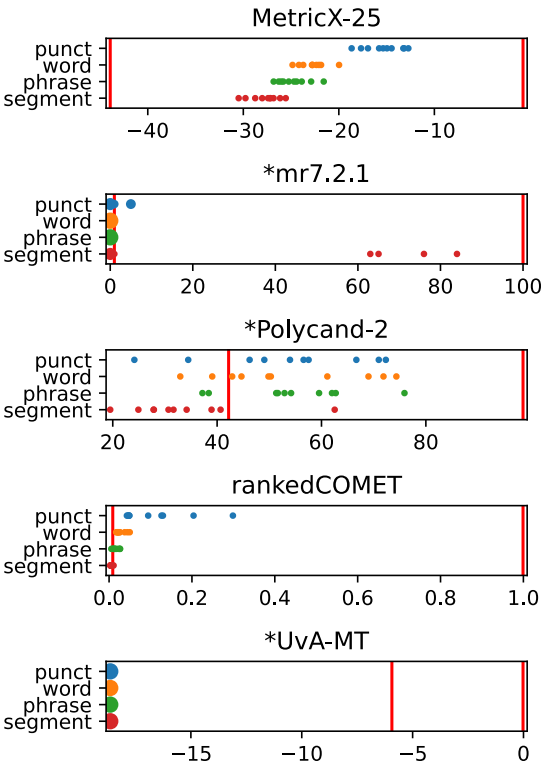

Figure 4: Japanese→Chinese scores assigned to empty string hypothesis paired with real (non-empty) source and reference. Where multiple strings receive the same score, this is indicated by proportionally increased dot size (in the case of **UvA-MT*, all strings received the same score, as indicated by the large dots). Red vertical lines indicate the minimum and maximum scores assigned over all Japanese→Chinese WMT News primary submission data. Asterisks indicate reference-free (QE) metrics.

rankedCOMET show a pattern that was observed in Knowles et al. (2024), where longer strings (segments) receive lower scores and some of the shorter strings (punctuation) receive higher scores, matching the intuition that a full sentence or document is more different from the empty string than a single character is. For both **Polycand-2* and **UvA-MT* we observe the opposite trend, with higher scores assigned to the longer strings. In both cases, most of the scores assigned to punctuation, words, and phrases are lower than the scores assigned to any of the submitted MT system data, but some of the scores assigned to the segments fall close to the middle of that score range. While both are reference-free metrics (i.e., not using the information that the reference string is the empty string), they do still have access to the source, making the result on segments more surprising.

The case of empty source and reference paired with non-empty hypothesis is an extreme representation of overgeneration, generating text that is not grounded in the source. We argue that unusual results on this set of data should raise questions for more exploration of a metric's performance on instances of overgeneration. It will require additional study to determine if there is a link between these, or if these results are confined to this particular corner case.

Figure 4 is the corresponding figure for the empty string hypothesis paired with a real (non-empty) source and reference. This is an extreme case of undergeneration (failing to generate any output). Two reference-free metrics, **mr7.2.1* and **UvA-MT*, assign very low scores to most of these examples, though **mr7.2.1* assigns scores in the top half of its score range to some segments. The results for *rankedCOMET* are quite similar to the results on the previous set of experiments, with slightly higher scores assigned to punctuation, but generally low scores overall. *MetricX-25* follows a similar pattern with the shorter strings scoring higher, but overall in the middle of the score range, while **Polycand-2* does not show a clear pattern.

Once again, we argue that assigning relatively high scores to empty string hypotheses may indicate that metrics are failing to pick up on undergeneration. Additionally, assigning non-lowest scores to the empty string presents a potential mismatch between metrics and typical standards for human evaluation (i.e., human annotators instructed to, or otherwise deciding on their own to, give the lowest scores to empty translations).

6.2 Mixed- and Wrong-Language Text

Metric	mix>wrong	equal	wrong>mix
**COMETKiwi22*	1	0	17
MetricX-25	2	0	16
**mr7.2.1*	18	0	0
**Polycand-2*	17	0	1
**UvA-MT*	1	0	17

Table 1: Comparison of scores of mixed and wrong-language text. Systems indicated with an asterisk (*) are reference-free (QE) metrics. All baseline reference-based metrics ranked all 18 mix>wrong.

As described in Section 3.2.2, we explore the scores that metrics assign to mixed- and wrong-language text. For our Japanese→Chinese challenge set for this task, the mixed language text is a mix of English and Chinese, intended to contain the full semantic information of the source. The wrong language text is English. Both the Chinese reference (used to build the basis of the mixed-language text) and the Japanese source are actually translations of the English data (which we also use to construct the mixed-language text). Due to the overlap between the Chinese reference and the mixed-language hypothesis, almost all baseline and primary reference-based metrics score the mixed-language hypothesis higher than the wrong-language for all 18 examples. The one exception to this is *MetricX-25*, as shown in Table 1, which scores the wrong-language text higher than the mixed-language text in 16 of the 18 examples. Both **COMETKiwi22* and **UvA-MT* score the wrong-language text above the mixed-language text in 17 out of 18 examples, while **Polycand-2* does the reverse and **mr7.2.1* prefers the mixed language text in all cases.

It remains an open question how mixed-language text should be scored, and is likely dependent on the intended audience of the translation. In any case, it may be surprising to observe systems preferring hypotheses that contain none of the intended target language at all over those that do at least include some target language text. This highlights—particularly with the shift to reference-free and multilingual metrics—the importance of taking into account the intended target language in evaluation. While we use a very small dataset here (18 segments), the consistency that we observe within metrics is notable. The issue of wrong-language output continues to be one that appears to be under-

examined by the designers of metrics, a claim we make based not only on this small-scale proof-of-concept, but by similar work in past challenge sets across more languages (Amrhein et al., 2022, 2023; Knowles et al., 2024).

6.3 English Language Spelling Variants

With the inclusion of an increasing amount of region information for the languages in WMT, we were interested in exploring English language spelling variants. As a preliminary step, we explored common British and American spelling differences. We used pairs of Czech→English segments where the English hypothesis varies only in the spelling conventions for certain terms. We submitted this portion of the challenge set three different times, once each with the language/region described as "en", "en_GB", and "en_US". We observed no difference in metric preferences depending on the choice of region descriptor; it is likely that most of these metrics are not taking into account the regional information at this granularity (compare also to the results in Section 6.2, which suggest that even the language code itself may not be entirely influential). For the three reference-free metrics that participated in this portion of the challenge set, we observed three different results (Table 2): **COMETKiwi22* was equally split between US and GB, but rarely scored them identically, **mr7.2.1* scored them identically more than half of the time and preferred GB almost half of the time, and **Polycand-2* scored the US variant higher the majority of the time. Due to the setup of the experiment, we could also check whether repeated instances of the same examples were scored identically; for **Polycand-2* there were some small (up to $3.81e-06$) differences in repeated scores; the differences between the US and GB variants were substantially larger.

We manually examined the error span results for baseline *XCOMET* and the two primary submissions of **AIP1* and *GemSpanEval* but did not observe any clear patterns related to the spelling variants. *GemSpanEval* did label some of the spelling variant terms as errors, but there was not a clear pattern related to the intended target language variants.

As the WMT shared tasks shift to include more region information, we expect that metrics will seek to handle this as well. We choose English for this particular example, because variation in English has been overlooked at WMT, even in instances when Englishes are paired with regionally-specified language variants. While we focused on English variation in the target; metric biases may also be relevant where the source is concerned.

Metric	US>GB	Equal	GB>US
**COMETKiwi22*	9	2	9
**mr7.2.1*	2	11	7
**Polycand-2*	16	0	4

Table 2: Comparison of reference-free (QE) metrics on pairs of Czech→English sentences where the English hypothesis only varies in whether certain terms use British or American English spelling conventions. For the 20 examples, the table shows the counts of those for which the US spelling version was given a higher score, for which the scores were equal, and for which the British spelling convention was given a higher score. Results are identical regardless of whether the intended target language has the region specified or not ("en", "en_GB", "en_US").

Both the dataset we used and the number of metrics that completed the task are much too small to draw broader conclusions from. Nevertheless, we think this will be an interesting avenue to explore, as WMT shifts to incorporate more regional information into its translation tasks.

7 Conclusions

We observe similar results to past MSLC experiments, with some metrics struggling to accurately score extremely low-quality (nonsensical) MT outputs. We continue to encourage discussion around how metrics should score empty strings and encourage additional analysis of how this does or does not correlate with broader metric sensitivity to over-generation and undergeneration. As we see more metrics shifting to use multilingual embeddings our large language models and more reference-free metrics, we encourage metric builders to consider how to incorporate information about the intended target language into their metrics (see also, Zouhar et al. (2024)). While it may be easy for a human—even one who cannot read the languages in question—to tell if an MT system has erroneously generated English when Chinese was expected, it may not be so simple for more similar language pairs. We would encourage metric builders to consider how to incorporate intended target language into their systems, and note that this may be an area where ignoring available references may have a real cost when it comes to metric trustworthiness.

Limitations

We focus on a small set of language pairs (in fact, smaller than in past iterations) and use small dataset sizes. This year, we did not submit systems to the General MT task, which means that we do not have a way to confirm how close those systems are (by human evaluation) to submitted systems; this may result in a gap in coverage between low and high quality systems. In general, these experiments represent corner cases that metrics builders should be considering in their systems. Our results primarily serve to flag issues to potential users of metrics and to encourage builders of metrics to test their metrics extensively.

Acknowledgements

We thank the WMT Metrics Task and WMT General MT Task organizers for permitting us access to the references in order to build this challenge set. We thank the reviewers for their comments and suggestions.

References

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2023. ACES: Translation accuracy challenge sets at WMT 2023. In *Proceedings of the Eighth Conference on Machine Translation*, pages 695–712, Singapore. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum Bayes risk decoding: A case study for COMET. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1125–1141, Online only. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.

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.

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.

Liane Guillou, Christian Hardmeier, Ekaterina Lapshinova-Koltunski, and Sharid Loáiciga. 2018. A pronoun test suite evaluation of the English–German MT systems at WMT 2018. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 570–577, Belgium, Brussels. Association for Computational Linguistics.

Michael Hanna and Ondřej Bojar. 2021. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online. Association for Computational Linguistics.

Felix Hieber, Michael Denkowski, Tobias Domhan, Barbara Darques Barros, Celina Dong Ye, Xing Niu, Cuong Hoang, Ke Tran, Benjamin Hsu, Maria Nadejde, Surafel Lakew, Prashant Mathur, Anna Currey, and Marcello Federico. 2022. Sockeye 3: Fast neural machine translation with pytorch. *arXiv*, abs/2207.05851.

Miroslav Hrabal, Ondrej Glembek, Aleš Tamchyna, Almut Silja Hildebrand, Alan Eckhard, Miroslav Štola, Sergio Penkale, Zuzana Šimečková, Ondřej Bojar, Alon Lavie, and Craig Stewart. 2025. Cuni and phrase at wmt25 mt evaluation task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. Metricx-25 and gemspaneval: Google translate submissions to the wmt25 evaluation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.

Rebecca Knowles, Samuel Larkin, and Chi-Kiu Lo. 2024. MSLC24: Further challenges for metrics on a wide landscape of translation quality. In *Proceedings of the Ninth Conference on Machine Translation*, pages 475–491, Miami, Florida, USA. Association for Computational Linguistics.

Samuel Larkin, Chi-Kiu Lo, and Rebecca Knowles. 2024. MSLC24 submissions to the general machine translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 139–146, Miami, Florida, USA. Association for Computational Linguistics.

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

Chi-kiu Lo, Rebecca Knowles, and Cyril Goutte. 2023a. Beyond correlation: Making sense of the score differences of new MT evaluation metrics. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 186–199, Macau SAR, China. Asia-Pacific Association for Machine Translation.

Chi-kiu Lo, Samuel Larkin, and Rebecca Knowles. 2023b. Metric score landscape challenge (MSLC23): Understanding metrics' performance on a wider landscape of translation quality. In *Proceedings of the Eighth Conference on Machine Translation*, pages 776–799, Singapore. Association for Computational Linguistics.

Sujal Maharjan and Astha Shrestha. 2025. Rankedcomet: Elevating a 2022 baseline to a top-5 finish in the wmt 2025 qe task. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. 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.

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, and 20 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

OpenAI. 2025. Openai o3 and o4-mini system card.

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

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

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

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

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, 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.

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The word sense disambiguation test suite at WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 588–596, Belgium, Brussels. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Di Wu and Christof Monz. 2025. Uva-mt at wmt25 evaluation task: Llm uncertainty as a proxy for translation quality. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

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.

Taemin Yeom, Yonghyun Ryu, Yoonjung Choi, and JinYeong Bak. 2025. Tagged span annotation for reasoning llm-based translation error span detection. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

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

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

Maike Züfle, Vilém Zouhar, Tu Anh Dinh, Felipe Maia Polo, Jan Niehues, and Mrinmaya Sachan. 2025. Comet-poly: Machine translation metric grounded in other candidates. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.