MSLC24: Further Challenges for Metrics on a Wide Landscape of Translation Quality

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 second edition of the Metric Score Landscape Challenge (MSLC), we examine how automatic metrics for machine translation perform on a wide variety of machine translation output, ranging from very low quality systems to the types of high-quality systems submitted to the General MT shared task at WMT. We also explore metric results on specific types of data, such as empty strings, wrong- or mixed-language text, and more. We raise several alarms about inconsistencies in metric scores, some of which can be resolved by increasingly explicit instructions for metric use, while others highlight technical flaws.

1 Introduction

This work builds on Lo et al. (2023), which introduced the Metric Score Landscape Challenge (MSLC).[1] At the Conference on Machine Translation (WMT), the Metrics Shared Task typically focuses on high-performing machine translation (MT) systems, in order to determine which new and improved metrics provide the most accurate and reliable scores (via comparison to human evaluation). However, the goal is for these metrics to go on to be used more broadly, which will likely result in their use on a wider range of systems. Since the Metrics Task primarily focuses on high-performing MT systems and their human evaluations, there is a risk that the new knowledge generated by the task about metrics may not generalize to lower-quality MT. For this reason, we submit a challenge set that covers a wider range of MT quality, in order to give potential users as well as metrics researchers a view of a broader range of performance. We also consider specific phenomena that may result in unexpected results from some metrics. We focus on three language pairs:English→Spanish (`eng`→`spa`), English→German (`eng`→`deu`), and Japanese→Chinese (`jpn`→`zho`).

2 Data

We divide this MSLC into two subsets: the first challenge set (MSLC-A) follows the approach set out in MSLC23, merging together our low- to mid-quality systems with the systems submitted to the General MT shared task, while the second challenge set focuses on specific phenomena (MSLC-B; developed based on notable results from 2023 and new aspects of this year's General MT Task).

2.1 MSLC-A: News Data

We focus only on the "news" subset of the WMT General Task test set, as this better matches the domain of our trained MSLC systems and because of concerns with some of the other domains. All figures and values for MSLC-A will be shown over the subset of the "news" data that was manually evaluated with MQM (Multidimensional Quality Metrics; Lommel et al., 2013) by the Metrics Shared Task unless otherwise noted.

The MSLC-A systems we evaluate are a range of low- to medium-quality sets of MT output for the three identified language pairs.

The MT models we build for MSLC24 are all constrained (as per the WMT General Task rules) models, built using standard WMT training data (or subsets thereof), without the application of common additional techniques like backtranslation or tagging. We train all NMT models using Sockeye version `3.1.31` (Hieber et al., 2022), commit `13c63be5` with PyTorch `1.13.1` (Paszke et al., 2019).

The English→German systems are the same ones described in Lo et al. (2023); we direct the reader to that work for more details. The English→Spanish and Japanese→Chinese systems are described in more detail in Larkin et al. (2024). We use checkpoints from training the systems as

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

Proceedings of the Ninth Conference on Machine Translation (WMT), pages 475–491
November 15–16, 2024. ©2024 Association for Computational Linguistics

representative of varying levels of quality. The levels of quality are manually checked by authors familiar with the relevant target languages on a small sample of the data. We list the checkpoints used for the systems in Appendix A. The lowest-quality systems are indicated with the letter A, and the quality approximately increases as the system labels proceed alphabetically.

2.2 MSLC-B: Specific Phenomena

We target three specific phenomena in the MSLC-B challenge set: empty strings, mixed- and wrong-language text, and language variants. In addition to this, across these, we consider an overarching theme of consistency. We begin by describing and justifying our study of these phenomena and the topic of consistency.

Lo et al. (2023) observed unusual performance around empty strings (which appeared due to a submitted system's output in 2023). This may, at first glance, seem like a trivial and uninteresting issue. We argue that it is worth exploring, for three primary reasons: it is a real scenario that we observe in the WMT submissions and in more general MT (empty strings *do* appear in output and sometimes even input or references), it is important to know how metrics handle the empty string (as different metrics take different approaches to handling empty strings), and because of the question of consistency (some metrics may score empty strings in internally-inconsistent or surprising ways). It would be simple for all implementations of metrics to treat empty strings (in the source, reference, or hypothesis) as an edge case to be handled separately; in practice this is not what we observe, so it is important for users of metrics to be aware of how metrics may perform in these cases.

We also consider questions of how metrics perform when the MT output is mixed-language or wrong-language text. This is a situation that can arise, for example, due to noise in training data.

In a similar vein, since the General MT Task specified translation into Latin American Spanish, we build a very small test set of terms that differ between variants of Spanish spoken in Latin America and in Spain. For example, the word *computer* may commonly be translated as *ordenador* in Spain but *computadora* in Latin America. We use this to examine how metrics, particularly reference-free metrics, score translations from different language variants. This is a very small-scale study, but our results indicate that this is an area that should be considered for future work.

We now describe how we build this portion of the challenge set in order to study these issues.

2.2.1 English→German and English→Spanish

Here we produce a small data set to explore these issues more closely. We begin by selecting data that will be used repeatedly:

- 10 segments (paragraphs and sentences) from the English language source (WMT news data) with their Spanish and German[2] reference translations
- 10 short phrases in English with reference translations (confirmed via wikipedia, Linguee, and WordReference)[3]
- 10 words in English with reference translations
- 10 punctuation marks or other characters

Taking all of these, we consider the following situations: empty source and reference paired with the reference segments described above (simulating an MT system generating fluent text after empty string input) and empty string hypothesis paired with the known source and reference (simulating an MT system outputting the empty string).

Using only the segment (paragraph or sentence length) portion, we also consider the situation where the output is fluent but in the wrong-language by pairing the source with the correct reference but the opposite language hypothesis (e.g., English source, Spanish reference, German reference used as hypothesis). We also consider a mixed-language hypothesis, manually produced by substituting substrings of the Spanish reference with substrings from the German `refB` reference.[4] For German, because we have access to `refB`, we also submit a version with English source, `refA` as the reference, and `refB` as the hypothesis; this permits a full range from incorrect language to

[2] For the German references we use `refA`.

[3] https://en.wikipedia.org/, https://www.linguee.com/, https://www.wordreference.com/

[4] This was done in such a way to maintain (approximate) fluency and adequacy, such that a reader familiar with both German and Spanish should still be able to understand the text. We would urge caution, however, in assuming that any results from this part of the test set would extend to more natural code-mixing.

Source	Last year, the World Economic Forum forecast that it would take five generations to achieve gender equality in every nation. Now the World Bank wants to rapidly accelerate that time frame.
Reference	Im vergangenen Jahr hat das Weltwirtschaftsforum vorausgesagt, es würde fünf Generationen dauern, bis in allen Staaten Geschlechtergleichstellung herrsche. Die Weltbank hat sich nun zum Ziel gesetzt, diesen Zeitraum deutlich zu verkürzen.
refB	Im vergangenen Jahr prognostizierte das Weltwirtschaftsforum, dass es fünf Generationen dauern werde, die Gleichstellung der Geschlechter in jeder Nation zu erzielen. Jetzt möchte die Weltbank diesen Zeitrahmen erheblich verkürzen.
Mixed-Lang.	*El año pasado,* prognostizierte das Weltwirtschaftsforum *que harían falta cinco* Generationen *para lograr la igualdad de género* in jeder Nation. Jetzt möchte die Weltbank *acelerar ese plazo rápidamente.*
Wrong Lang.	El año pasado, el Foro Económico Mundial pronosticó que harían falta cinco generaciones para lograr la igualdad de género en todas las naciones. Ahora, el Banco Mundial quiere acelerar ese plazo rápidamente.

Table 1: Example of wrong-language, mixed-language (Spanish shown in italics), and refB (correct language alternate human reference) as hypotheses in the English→German MSLC-B dataset.

mixed-language to matched language (but different human translation). We show an example of this in Table 1.

For Spanish, since the WMT General MT Task explicitly describes this translation task as "EN to Spanish (Latin America)", we provide a very small sample (8 words) of words that tend to have differing translations between varieties of Spanish spoken in Latin America and varieties of Spanish spoken in Spain. This has very limited coverage but may permit us to begin asking questions about whether quality estimation systems have tendencies or biases towards certain language varieties.

2.2.2 Japanese→Chinese

For Japanese→Chinese, we examine metrics' performance around empty strings by first selecting data that will be used repeatedly:

- 5 segments (paragraphs and sentences) from the Japanese language source with their Chinese reference translations
- 5 short phrases in Japanese with reference translations
- 5 words in Japanese with reference translations
- 5 punctuation marks in Japanese with reference translations

We consider the same two types of empty string situations as in the other language pairs. The empty strings challenge examples make up 40 items in the MSLC-B Japanese→Chinese test set.

Similarly to the other language pairs, we consider wrong-language output (an English translation of the Japanese source, produced as a human translation from the Chinese reference by one of the authors) and mixed-language output (substituting words or phrases in the Chinese reference with corresponding Japanese and English words or phrases); these make up 10 items in the MSLC-B Japanese→Chinese test set.

3 Metrics

There are dozens of metrics submitted by the task organizers and participants to the WMT24 Metrics Shared Task. Given time and space limitations, we only examine the baseline metrics submitted by the task organizers and the primary metrics submitted by the participants. We describe the metrics included in this work in Appendix B.

4 Results and Plots

We divide our examination of the results into the two parts of the challenge set: MSLC-A and MSLC-B.

4.1 MSLC-A

Here we present preliminary results for the MSLC-A subset of the challenge set. We begin with the segment level and then consider system-level results. We make use of the MQM results provided by the Metrics Task organizers.

4.1.1 Segment Level

The histograms along the diagonal of Figure 1 show the distributions of segment-level scores produced by a subset of the baselines and submitted primary metrics. We can see that different metrics exhibit very different score distributions. Some show a somewhat bimodal distribution of scores, others are closer to normally distributed. For the metrics that are closer to normally distributed, we also see different skews. Most metrics are left skewed (i.e., they more frequently give segment scores in the higher-end of their possible score range), while BLEU is right skewed and more frequently gives segment scores in the lower end of its possible score range.

Metrics also differ in whether they exhibit a strong separation between the segments produced by the low-quality systems from our challenge set and the segments produced by the WMT submissions or whether they assign a range of low to high scores to most systems (i.e., having clear overlap in score range across all systems). This variation in characteristics suggests that metrics may have different strengths and weaknesses across the translation quality landscape; not all metrics are equally appropriate for scoring high-quality and low-quality MT.

XCOMET gives very low scores to segments from the very low-quality systems, but uses much more of the score space for the mid-quality systems. On the low-quality side, this is somewhat similar to the distribution of *BLEU* scores, but the high-quality systems have *XCOMET* scores that are much higher due to *XCOMET*'s bimodal distribution. Meanwhile, *chrF* shows a fairly normal distribution, but with a clear distinction between the various MSLC systems. We can also see this reflected when we examine system-level scores.

There are also metrics that use an approximation of a discrete score space, such as *GEMBA-ESA*. Lo et al. (2023) noted several metrics that did this in 2023; *GEMBA-ESA* is the only one in this year's set that does.

4.1.2 System Level

To analyze system-level scores, we compute an average over all of the segment-level scores in the news domain for a given MT system. There are two reasons why we are using this segment average instead of the submitted system-level score: 1) not all metrics submitted system-level scores and 2) using averaged segment-level scores allow us to show a representation of uncertainty (computed with bootstrap resampling, 1000 times, for $p < 0.05$) for the metrics. These system-level scores can also be used in order to gain a better understanding of the overall range of a metric's scores, as well as what kind of scores are assigned to very low quality machine translation (e.g., the A and B systems from the challenge set).

Figure 2 shows the system average scores for a subset of English→German (see Appendix D.1 for other translation directions). We observe that metrics show different patterns of scores at the system level. Both *PrismRefMedium* and *PrismRefSmall* appear to have serious difficulties in accurately scoring the lowest-quality system and give it a score higher than some of the better (still low-quality) systems.[5] Some metrics, such as *GEMBA-ESA*, *XCOMET* and *XCOMET-QE*, give very close scores to all of the low-scoring systems. For a use case (e.g., a low-resource language) where one expects to have low- to medium-quality systems at least initially, one may want to choose a metric that provides clearer distinctions between various systems on the lower range of quality.

For the high-quality systems the string-based metrics, such as *BLEU* and *chrF*, show wider error bars and thus may not distinguish between them. We leave analysis of the high-quality systems to the Metrics Shared Task.

By having our top MSLC system evaluated alongside the submitted WMT systems, we are able to observe that for Japanese→Chinese our systems combined with the high-performing submitted WMT systems do cover the wide range of quality. For English→German and English→Spanish, however, there may be a "missing middle" gap in quality that is not covered, an issue we aim to address in future work.

4.1.3 Conclusions: MSLC-A

As we saw in Lo et al. (2023), metrics differ in how they use their available score space. Some make fairly full use of their score range, others discretize the score space, and yet others display bimodal distributions of scores. All of these impact how individual segments are scored as well as how the system-level scores are distributed (i.e., whether the system-level scores are distributed more uniformly over the score space from low quality to

[5] Though we only do small-scale informal human evaluation, we expect, e.g., system E should not be ranked below system A.

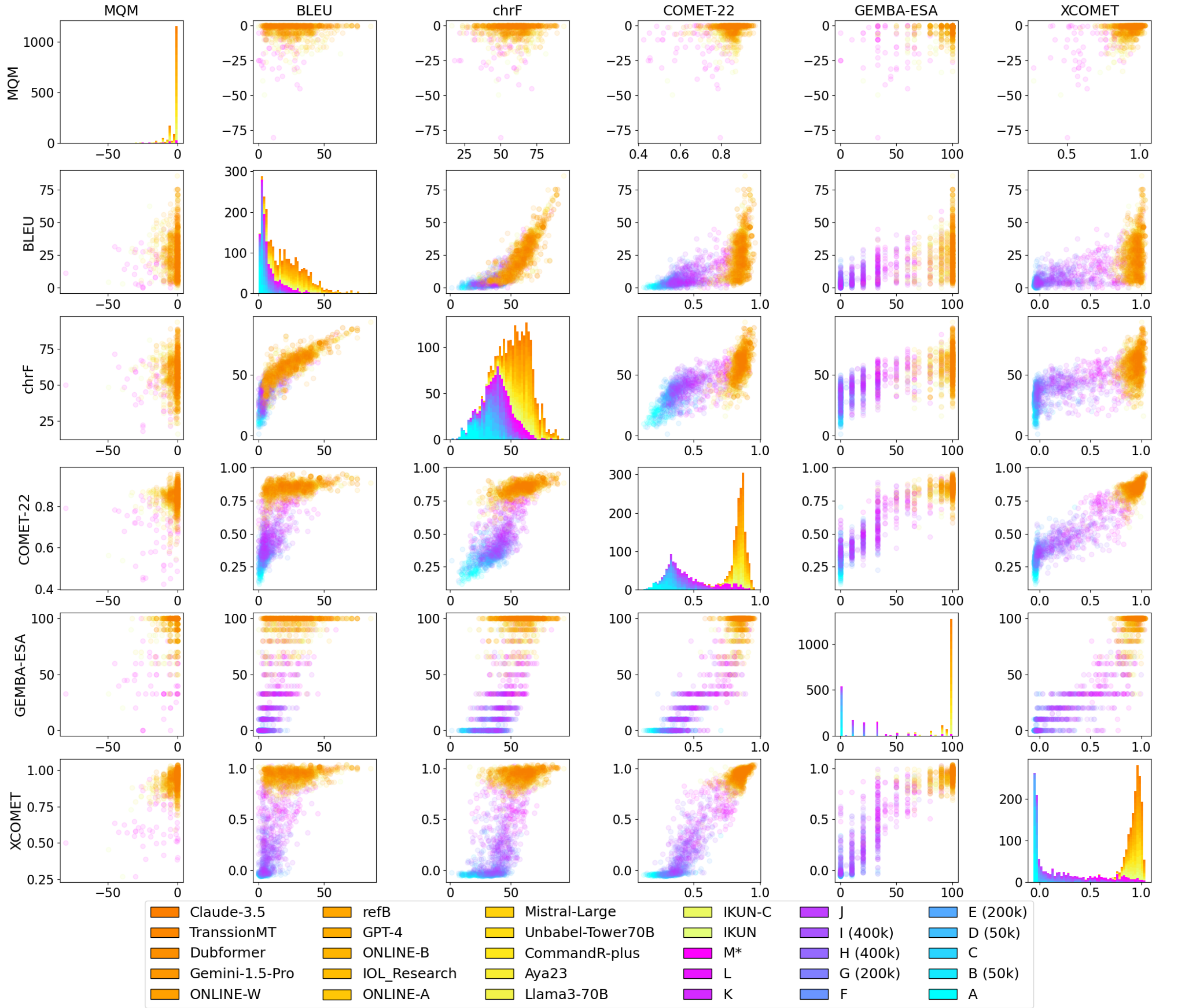

Figure 1: Matrix of segment-level scores for English→German. Along the diagonal are stacked histograms of segment scores across the challenge set (cool colours/bottom) and submitted WMT systems (warm colours/top). The off-diagonal entries are scatterplots where each point is a single segment positioned according to the score assigned to it by row and column metrics; each point is coloured according to the same colours as the histogram. Note: for a full, scalable version of this figure, see `https://github.com/nrc-cnrc/MSLC`; all other figures in this paper are scalable.

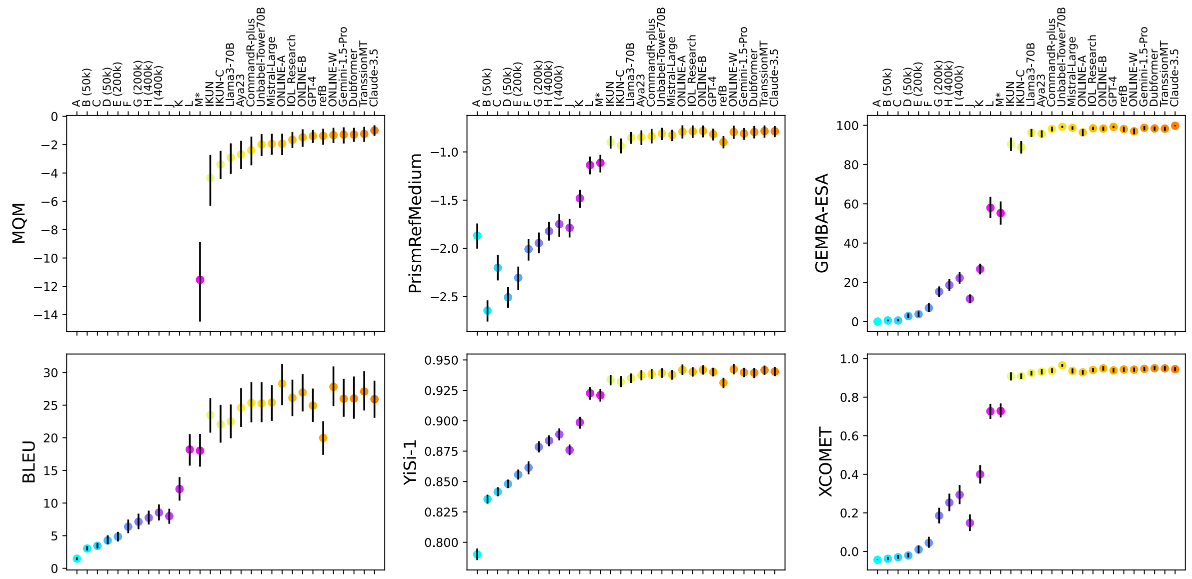

Figure 2: System average scores for English→German. MSLC systems (cool colours, left) are ordered by BLEU score and brief manual examination; WMT submitted systems are ranked by average MQM score.

high quality, or whether most systems are clustered near the low and high ends of the score space). This year we noted fewer extremely unusual distributions; we did not see a repeat of the "universal scores" results observed in Lo et al. (2023).

The MQM evaluation of our top-performing system helps us to get a better idea of how to interpret these scores, though we note the issue of the missing middle range of human scores in two out of our three language pairs. We also note a weakness of error-based evaluations: they may not always capture non-errors (e.g., ways of translating that are not incorrect, but may be dispreferred by translators or end-users).

In future work, we may wish to apply more formal human evaluation to our lower-scoring systems, to better clarify the full range, but this year's introduction of human scores for one system per language pair takes a step towards that goal.

4.2 MSLC-B: Empty Strings

In MSLC23 (Lo et al., 2023), we observed a variety of system scores on empty strings produced by one of the participating systems in the WMT task. Here, we expand on that in a controlled fashion, examining the scores that metrics output when scoring empty strings.

4.2.1 Empty Source and Reference

We begin with empty source and reference, paired with four different types of output: single punctuation characters (*punct*), single words (*word*), short phrases (*phrase*), and full sentences/short paragraphs (*sent*). All of the hypothesis text is in the target language, with the full sentences drawn from the WMT news data reference (`refA`, in the case of German). If these had been produced by an MT system taking an empty source and generating text, this might be considered a "hallucination"—generating fluent text that is not conditioned on any relevant source text. As such, we would expect that MT metrics should give these low scores. While some metrics (*BERTScore*, *BLEU*, *YiSi-1*, *chrF*, *spBLEU*, *mmm_qe*) do consistently give their lowest score (0) to all of these test segments, others show a greater variety of results.

Figure 3 shows a subset of the remaining metrics for English→German, covering a range of the variations in scores. Each subfigure shows the scores assigned to the 10 items in each category, with the vertical red lines indicating the lowest and highest scores assigned by this metric to any of the WMT news test data for any submitted MT system. *COMET-22* demonstrates the most common pattern: assigning a range of scores, with a tendency to have slightly higher scores for the shorter categories (e.g., punctuation—a single character has a very small edit distance to the empty string, perhaps making it more similar to the empty string than longer text) and lower scores to items in the longer categories (i.e., penalizing generating a full sentence out of nothing). *PrismRefMedium* and

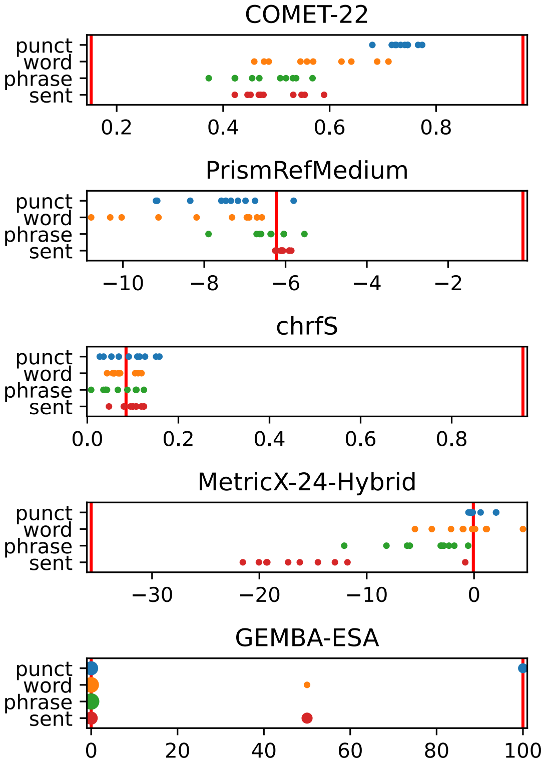

Figure 3: English→German scores assigned to text when paired with empty source and reference. Red vertical lines indicate the minimum and maximum scores assigned over all WMT News primary submission data.

chrfS show another common pattern, by assigning low scores to all items; this is more in line with the desired and anticipated performance on this set of data. We note that the scores from *chrfS* are quite clustered around the lowest scores assigned to the WMT news data, while *PrismRefMedium* has scores expanding into a much lower range than the range of scores it assigned to the news data in the WMT test set. *MetricX-24-Hybrid* shows concerning results on this test set, assigning scores *higher* than any assigned to the WMT news test data to some of the samples, particularly the punctuation (perhaps not entirely unreasonably, as MT systems may need to occasionally generate additional punctuation in the target language), but also in some word and phrase examples. Finally, *GEMBA-ESA* assigns its lowest score most of the time, but occasionally assigns a top score or a score exactly in the middle of the range, an unexpected inconsistency.

4.2.2 Empty Hypothesis

Next, we flip the empty strings to the output side and pair them with real sources and references for the four different types of text mentioned in the previous subsection. This is simulating the extreme case of omission where the complete output is missing. We understand that MT users may find it acceptable to omit translation for a single punctuation. As such, we again may expect that MT metrics would give gradually lower scores to the empty string output as the length of the source and reference increase. Similarly to the empty source and reference test cases, some metrics (*BERTScore*, *BLEU*, *YiSi-1*, *chrF*, *spBLEU*, *mmm_qe*) do consistently give their lowest score (0) to all of these test segments.

Figure 4 shows a subset of metrics for English→German, covering a range of the variations in scores. As we observe, *PrismRefMedium* and *chrfS* also give low scores (although not their lowest possible score) to empty string output. Some metrics (e.g. *COMET-22*) indeed give gradually lower scores to the empty string output according to the length of the input, with the items in the *sent* category receiving the low end of scores. We find that this is still relatively unsurprising behavior for metrics scoring empty string output. However, *MetricX-24-Hybrid* and *XCOMET* show concerning results on this test set, assigning mid-range to high scores to empty string output. Finally, as was the case in the empty source and reference test set, *GEMBA-ESA* assigns its lowest score most of the time, but occasionally assigns a top score to the empty string output.

4.2.3 Conclusions: Empty Strings

These empty string test cases (both empty source and reference and empty output) reveal undesirable metric results: giving high scores to extreme hallucination and omission. This leads us to be particularly concerned about the decision by the WMT General MT Task to use *MetricX-23-XL* and *CometKiwi-DA-XL* to decide which participating systems would receive human annotations, because related metrics (*MetricX-24-Hybrid* and *CometKiwi*) are two of the metrics showing these undesirable phenomena.

This may be an opening for a wider discussion about whether it is better for an MT system to fail to generate output than to generate output that is incorrect; nevertheless this would be a departure from past expectations (where, e.g., in human evaluation, “no translation” is typically given as a prototypical example of something that should receive a low score). In any case, we can likely find common ground in agreeing that metrics should not give

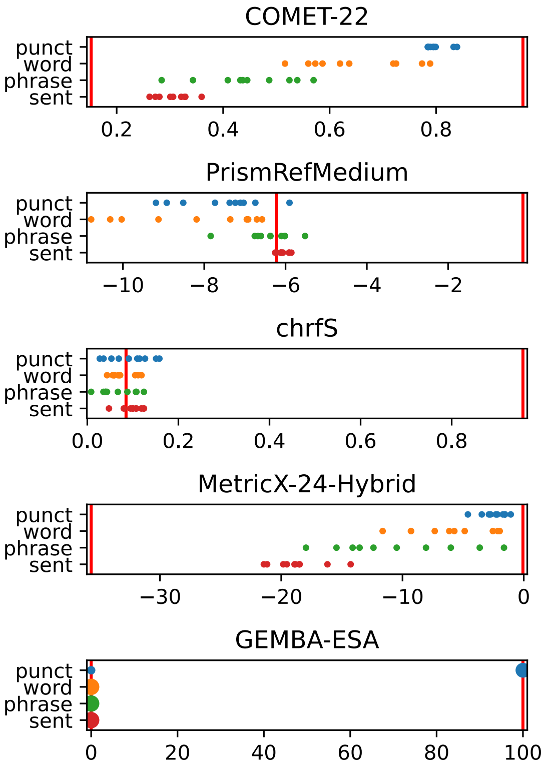

Figure 4: English→German scores assigned to the empty string paired with real source and reference.

high scores to non-empty output when given an empty input and empty reference. We would encourage a broader conversation about this, and in the meantime would encourage those presenting new metrics to be sure to specify how their metrics handle empty strings.

We encourage both metric builders and metric users to be aware of how metrics treat these edge cases. They do occur in practice and a user anticipating one type of performance on empty strings (e.g., low or 0 scores) may come to erroneous conclusions if they unknowingly use a metric that treats empty strings in another way (e.g., as high-scoring). We were also somewhat surprised to encounter the level of variation across empty string scores, and expect that users who are most familiar with string-matching metrics like BLEU may also not expect this variety of results.

4.3 MSLC-B: Mixed- or Wrong-Language

In this section, we explore what kind of output the metrics produce when they are applied to mixed-language and wrong-language translation hypotheses. We focus on English→German, because access to a second human-translated reference (`refB`), allows us to explore a range of translation hypotheses, from a good human translation (`refB`, in German), a mixed-language translation (composed of a mix of the text of `refB` and the human-translated Spanish reference), and a wrong-language translation (the Spanish language reference). Our small test set for this is composed of 10 English source sentences, along with the various translations described above. We would expect a well-performing and usable metric to assign high scores to the `refB` German translation, lower scores to the mixed-language translation (portions of which are correct German translations of words and phrases), and even lower scores for the Spanish translation (a fluent and accurate translation, but in the wrong language).

However, this is not precisely what we observe in Table 2. Most metrics do give a score to `refB` that is greater than or equal to the score given to the mixed-language text in all 10 examples, while others score it at or above the mixed-language the majority of the time (*mmm_qe* (9), *CometKiwi* (8), and *damonmonli* (6)). Only *XLsimMqm* scores the mixed-language text higher than `refB` in 8 out of 10 examples. When it comes to the wrong-language text, most metrics again score `refB` equal or higher all of the time, but others at least occasionally rank the wrong-language text above `refB`—reference-free metrics in particular tend to make some errors (with the exceptions of *GEMBA-ESA* and *MetricX-24-Hybrid-QE*), but the two *PrismRef** metrics also make these errors. When comparing the scores given to the mixed-language text and the wrong-language text, we see even more of a mix. Some systems (both *PrismRef** systems, *XLsimMqm*, *mmm_qe*, and *MetricX-24-Hybrid-QE*) never score the mixed-language text above the wrong-language text.

4.3.1 Conclusions: Mixed- and Wrong-Language

This varies somewhat between language pairs (see Appendix C), but string-based metrics like *BLEU* and *chrF* consistently score the mixed-language text above the wrong-language text. The weaknesses of string-based methods, such as their reliance on exact matches and lack of partial credit for synonyms (especially when evaluated against a single reference), have resulted in a shift towards embedding-based metrics that can provide more flexible semantic representations. However, given these results, it raises the question: are all modern

Metric	refB≥Mix	refB≥Wrong	Mix≥Wrong
BERTScore	10	10	8
BLEU	10	10	10
BLEURT-20	10	10	2
COMET-22	10	10	3
CometKiwi	8	4	1
PrismRefMedium	10	6	0
PrismRefSmall	10	7	0
YiSi-1	10	10	3
chrF	10	10	10
spBLEU	10	10	9
chrfS	10	10	10
MEE4	10	10	10
XLsimMqm	2	1	0
mmm_qe	9	6	0
mmm_hybrid	10	10	1
MetricX-24-Hybrid	10	10	2
MetricX-24-Hybrid-QE	10	10	0
GEMBA-ESA	10	10	10
XCOMET	10	10	1
XCOMET-QE	10	6	2
damonmonli	6	7	7

Table 2: `eng`→`deu`: Number of times (out of 10) that the metric scored `refB` higher than or equal to its mixed-language pair (refB≥Mix), higher than or equal to its wrong-language pair (refB≥Wrong), and a mixed-language hypothesis higher than or equal to its wrong-language pair (Mix≥Wrong).

metrics suitable for providing information about whether a text is a good translation *into the target language*, or simply whether it is a good translation (into some language(s))? We argue that these preliminary, small-scale results suggest the importance of additional analysis of this question. While this is unlikely to be a problem in many cases (especially when, e.g., language ID could also be performed), this may be particularly risky in low-resource settings where high-quality language ID is not available (cf. issues described in Kreutzer et al., 2022). In concurrent work, Zouhar et al. (2024) propose incorporating language ID to handle this issue as they explore it specifically in the context of *COMET*.

4.4 MSLC-B: Language Variants

The WMT General Task specifically called on researchers to build MT systems for English→Spanish using *Latin American* Spanish. We choose a small selection of terms that exhibit some of the vocabulary differences between the language variants of Spanish spoken in Latin America and Spain. We note several limitations to this: this is a very small set of terms, the terms are evaluated in isolation, and they are certainly not fully representative of all Spanish language variants spoken in Latin America or Spain.[6]

Due to the structure of the challenge set submission process, each source term was submitted four times: once for each language variant with the matching reference and once for each language variant with the opposite reference. Considering only the reference-free metrics (those that do not use the provided reference in order to compute their score), we observe results in Table 3. A checkmark(✓) indicates that in all cases for that term, the Latin American term chosen was scored higher than the term used more commonly in Spain; an ✗ indicates that the term used in Spain scored equal to or higher than the term used in Latin America. This somewhat arbitrary choice to include repeated versions was fortuitous, because it highlights a concern with one of the metrics: a question mark (**?**) in a cell indicates that the rankings computed were mixed. This means that, on repeated scoring, the variations within the metric scores returned were great enough (in the case of *MetricX-24-Hybrid-QE*) to result in different rankings at least once. This should be alarming to potential users of metrics, who would expect consistent results on repeated strings. That is: a user may reasonably expect that if they submit the same input to a metric twice in a row, they will get the same output twice in a row; here we observe that not all metrics have this as a guarantee. We discuss this more in Section 5. We note that other metrics also exhibited some variation in their scores, but the rest did not vary enough to change which of the two term variants received the higher score.

We now observe that *XLsimMqm* is the only metric to prefer the term used more commonly in Latin America more than half the time (5/8). We note that *GEMBA-ESA* only prefers the term from Latin America in 1/8 terms, but for the remaining 7 terms, both variants are given identical scores of 100 (*GEMBA-ESA* is the only metric whose counts would change, were we to score it as correct if a term used in Latin America scores *equal or greater than* the term used more commonly in Spain).

This raises similar questions to those we considered in the wrong-language and mixed-language experiments, albeit at a finer-grained level. Metrics may not be equally appropriate for use across all language variants, and may in fact demonstrate a scoring preference to one or the other. This will require considerably more experimentation, with larger test sets, in the future.

5 Consistency

Our experiments in MSLC-B highlighted some issues in metric score consistency: repeated instances of scoring the same string resulting in dif-

[6] In several of the cases presented here, there exist a number of other translations that we could have selected.

English source:	computer	sandwich	potato	juice	waiter	tires	peanut butter	drive	
Latin America:	computadora	sándwich	papa	jugo	mesero	llantas	mantequilla de maní	manejar	
Metric *Spain:*	ordenador	bocadillo	patata	zumo	camarero	neumáticos	crema de cacahuete	conducir	**Counts**
CometKiwi	✓	✓	✗	✓	✓	✗	✗	✗	4/8
XLsimMqm	✗	✓	✓	✓	✗	✗	✓	✓	5/8
mmm_qe	✓	✓	✗	✓	✗	✗	✓	✗	4/8
MetricX-24-Hybrid-QE	✗	✓	✓	✓	✗	✗	**?**	✓	4/8
GEMBA-ESA	✗	✓	✗	✗	✗	✗	✗	✗	1/8
XCOMET-QE	✓	✗	✗	✓	✗	✗	✗	✗	2/8

Table 3: Metric preferences for terms that are more common in Spanish language variants spoke in Latin America (✓), for terms more common to language variants spoken in Spain (✗), or inconsistent preferences (**?**).

ferent scores. The MSLC-B dataset provided only a small set of examples on which to test this, and only for reference-free metrics. However, because we submitted our highest-scoring MSLC MT systems to the WMT General MT task, we actually do have a larger set of data with which to explore repeated scores. This consists of the intersection between the MSLC-A data (news domain only) and the full General MT test set (149 segments for eng→deu and eng→spa and 269 for jpn→zho), each of which was scored by each metric as part of the MSLC-A challenge set and as a WMT submitted system.

It is important to highlight that, while the Metrics Shared Task calls for metrics that provide scores at the segment level and the system level, it does not currently specify how or when metrics may make use of extrasentential information (e.g., information from other parts of the test set or document) in order to produce segment level scores. This could include approaches that compute some statistics from the full test set (like *YiSi* does for TF-IDF) or that operate on the batch level (like *PrismRef**). This is to say, some of these apparent inconsistencies may be intentional (i.e., giving a segment a different score depending on the context in which it appears).

Since metrics have different score ranges, we first calculate the lowest and highest scores assigned by each metric to any of the MSLC-A or MSLC WMT submission segments within the news subset. This gives us a range of metric scores. Then, for each source-reference-hypothesis in the news subset, we compute the absolute difference in the score that it was assigned as part of the MSLC-A dataset and the score it was assigned as part of the MSLC WMT submission dataset and express this as a percentage of the metric's score range described above.

For many metrics (*BERTScore*, *BLEU*, *chrF*, *spBLEU*, *chrfS*, *XLsimMqm*, *mmm_qe*, *mmm_hybrid*, *GEMBA-ESA*), there is never any difference in these two scores. For other metrics, like *CometKiwi*, there are some small differences (never greater than 0.1% of the metric's score range); these seem likely attributable to rounding/floating point errors. In other cases, it is possible that other even larger differences may be accounted for due to differences in batch size and hardware used, such as the case of *MetricX-24-Hybrid-QE*, which sees its largest score difference as 7.3% of the score range for eng→deu.

For *YiSi-1*, there is a known reason for the observed differences (up to 2.3% eng→deu, 2.8% eng→spa, 4.4% jpn→zho): the *YiSi* score is computed taking into account TF-IDF statistics from the full test set; since MSLC-A included only the news data while the full WMT General Task submission for MSLC included other domains, the scores assigned to individual segments may differ, as the segment-level scores are conditioned on the full test set. The largest difference we observe is for *PrismRefMedium*, with one score difference of 98.9% of the full score range; this is likely also due to the model operating at the level of the document or document chunk. The MSLC-A challenge set did not include document boundaries, which could account for the differences we observe. In future tasks, we would suggest incorporating document information in the challenge set submission in order to avoid these issues, and it would also be helpful to clarify which metrics incorporate extrasentential information (specifically from other parts of the challenge set data). We know that there are at least three different levels on which metrics are operating: the single-segment level (i.e., each segment is scored individually, so repeated segments should be scored identically), incorporating information from the full test set (in which case repeated segments within the same test set may receive identical scores), and incorporating document/batch/multi-segment input (in which case, scores may depend on how the batching is performed). It could also be possible to have more complex interactions (e.g., taking into account where in a document a segment occurs in order to score it); metrics users and challenge set builders need to be aware of these in order to ensure that they are measuring what they think they are measuring.

As we can see, there are at least two different reasons for these apparent inconsistencies: 1) purposeful differences that arise from metrics that use contextual information for computing sentence-level scores (as in the case of *YiSi*) and 2) errors and noise resulting from computational or implementational factors. In the case of these purposeful differences, the primary thing for metric users to be aware of is the scope of the context that is used, in order to be able to reproduce scores. The latter issue is a larger problem, especially when we see score differences that cover substantial portions of the metric's score range. If a metric is unstable or produces different scores based on the hardware used to compute it, we face an issue at least as concerning as the preprocessing one identified in Post (2018). We propose two main (but not entirely satisfactory) solutions to this: 1) it may be best to report such metrics as an average over multiple runs and 2) metrics should adopt the proposals outlined in Zouhar et al. (2024) to include metric signatures for better reproducibility.

6 Conclusion

We once again show the diversity of ways that metrics perform on a wide range of system quality. We also observe quite a bit of variation in terms of how systems handle empty strings, which may influence how they are used (e.g., when comparing a system that frequently generates empty strings to one that never does). We also consider questions of wrong-language text and mixed-language text as well as language variants, and argue that metrics researchers should consider whether their metrics are overgeneralizing (i.e., whether they give high scores to good translations regardless of whether the translation is in the desired target language or not) or are biased towards particular language variants. Many of our results support the conclusions that Zouhar et al. (2024) describe in their concurrent analysis of *COMET*, such as the need to better handle empty strings, questions of target language, biases, and the importance of metric signatures when metric variations may introduce score differences. In concert with that work, we raise the concern that as new metrics are introduced, we are not learning the lessons from our field's past errors. We argue for the importance of examining real-world corner cases and issues of reproducibility in order to more responsibly introduce new metrics to the research community. Both metrics researchers and users should be alarmed by the levels of inconsistency that we observe. One of the benefits of using automatic metrics should be to make fair comparisons (for repeated scoring, across papers, and so forth)—inconsistent metrics cannot serve this purpose. When there are intentional sources of differences in scores for repeated segments (i.e., due to the context in which they appear), users need to be aware of the scope and approaches used to incorporate context, in order to ensure that they are using metrics as intended in order to measure what they intend to measure. This will become increasingly important as we see a shift to document-level

Limitations

We focus only on three language pairs (English→German, English→Spanish, and Japanese→Chinese) in the News domain in this work, due to the availability of human-annotated scores for this set. Several of our additional experiments use extremely small sets of data (e.g., 5-10 examples); in most cases these are designed to help us establish whether additional future study would be helpful, rather than to make definitive claims about the results. In time for the camera-ready submission, we had access to MQM scores, but not to the General MT Task ESA annotations.

Acknowledgements

We thank the WMT General MT Task organizers for permitting us access to the references in order to build this challenge set. We thank the WMT Metrics Task organizers for including our systems in their MQM human evaluation, to enable us to use those results to better understand the performance of automatic metrics across a range of MT quality.

References

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*, USA. 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.

J. D. Hunter. 2007. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

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*. Association for Computational Linguistics.

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

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using comet.

A MSLC24 MT Systems

In Table 4 we see the checkpoint IDs for systems included in the challenge set for eng→deu. Table 5 and 6 show the same for eng→spa and jpn→zho.

System	Checkpoints	*BLEU*
A	54	0.50
B (50k)	1	1.85
C	79	3.13
D (50k)	7	4.19
E (200k)	2	4.54
F	91	6.88
G (200k)	27	7.87
H (400k)	4	8.73
I (400k)	43	9.64
J	102	9.24
K	129	13.91
L	313	22.79
M (MSLC)	311	22.65

Table 4: Checkpoint IDs and *BLEU* scores (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1) on MSLC-A for systems included in challenge set (eng→deu); parenthetical numbers indicate one of the pseudo-low-resource systems rather than the full training data system.

System	Checkpoints	*BLEU*
A	52	0.75
B	65	4.94
C	74	8.55
D	84	13.14
E	98	19.91
F	123	25.61
G	207	31.47
H (MSLC)	800	37.97

Table 5: Checkpoint IDs and *BLEU* scores (nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1) on MSLC-A for systems included in challenge set (eng→spa).

System	Checkpoints	*BLEU*
A	37	0.05
B	70	4.97
C	80	10.74
D	97	15.79
E	133	19.48
F (MSLC)	358	23.12

Table 6: Checkpoint IDs and *BLEU* scores (`nrefs:1|case:mixed|eff:no|tok:zh|smooth:exp|version:2.3.1`) on MSLC-A for systems included in challenge set (`jpn`→`zho`).

Metric Name	**Reference-based**
Human annotation	
MQM	
Metrics	
BERTScore	✓
BLEU	✓
BLEURT-20	✓
chrF	✓
chrfS	✓
COMET-22	✓
CometKiwi	
damonmonli	✓
GEMBA-ESA	
MetaMetrics-MT	✓
MetaMetrics-MT-QE	
MEE4	✓
MetricX-24-Hybrid	✓
MetricX-24-Hybrid-QE	
prismRefMedium	✓
prismRefSmall	✓
sentinel-cand-mqm	
sentinel-ref-mqm	✓
sentinel-src-mqm	
spBLEU (flores-200)	✓
XCOMET	✓
XCOMET-QE	
XLsimMqm	
YiSi-1	✓

Table 7: Human annotation and metrics included in this work, with their coverage of language pairs. Metrics that are not marked as reference-based are reference-free (a.k.a quality estimation) metrics.

B Metrics

Table 7 shows a summary of the human annotations and metrics included in this work and the translation directions they participated in. For detail descriptions of the metrics, please refer to the Metrics Task overview paper (Freitag et al., 2024).

Note: in the main body of the text, for space reasons, we abbreviate the *MetaMetrics-MT-QE* and *MetaMetrics-MT* names as *mmm_qe* and *mmm_hybrid*, respectively.

Metric	**Mix≥Wrong**
BERTScore	10
BLEU	10
BLEURT-20	5
COMET-22	4
CometKiwi	3
PrismRefMedium	9
PrismRefSmall	8
YiSi-1	10
chrF	10
spBLEU	10
chrfS	10
MEE4	10
XLsimMqm	7
MetaMetrics-MT-QE	2
MetaMetrics-MT	1
MetricX-24-Hybrid	0
MetricX-24-Hybrid-QE	0
GEMBA-ESA	10
XCOMET	1
XCOMET-QE	0
damonmonli	6

Table 8: `eng`→`spa`: Number of times (out of 10) that the metric scored a mixed-language hypothesis higher than or equal to its wrong-language pair.

C Additional Mixed/wrong-language Tables

Tables 8 and 9 show the how the metrics scores mixed-language and wrong-language data for English→Spanish and Japanese→Chinese. For English→Spanish, the wrong-language text was German and the mixed-language was a mix of German and Spanish. For Japanese→Chinese, the mixed-language was a mix of Chinese, English and Japanese, while the wrong-language was English. Note that because the Chinese text in the mixed-language hypotheses is drawn directly from the reference, this should be a particularly easy task for string-based metrics.

D Additional Figures

Figures in this paper are produced using Matplotlib (Hunter, 2007), version 3.7.1.

D.1 MSLC-A System-Level

Figures 5, 6, and 7 show the system average scores for English→German, English→Spanish, and Japanese→Chinese across all metrics.

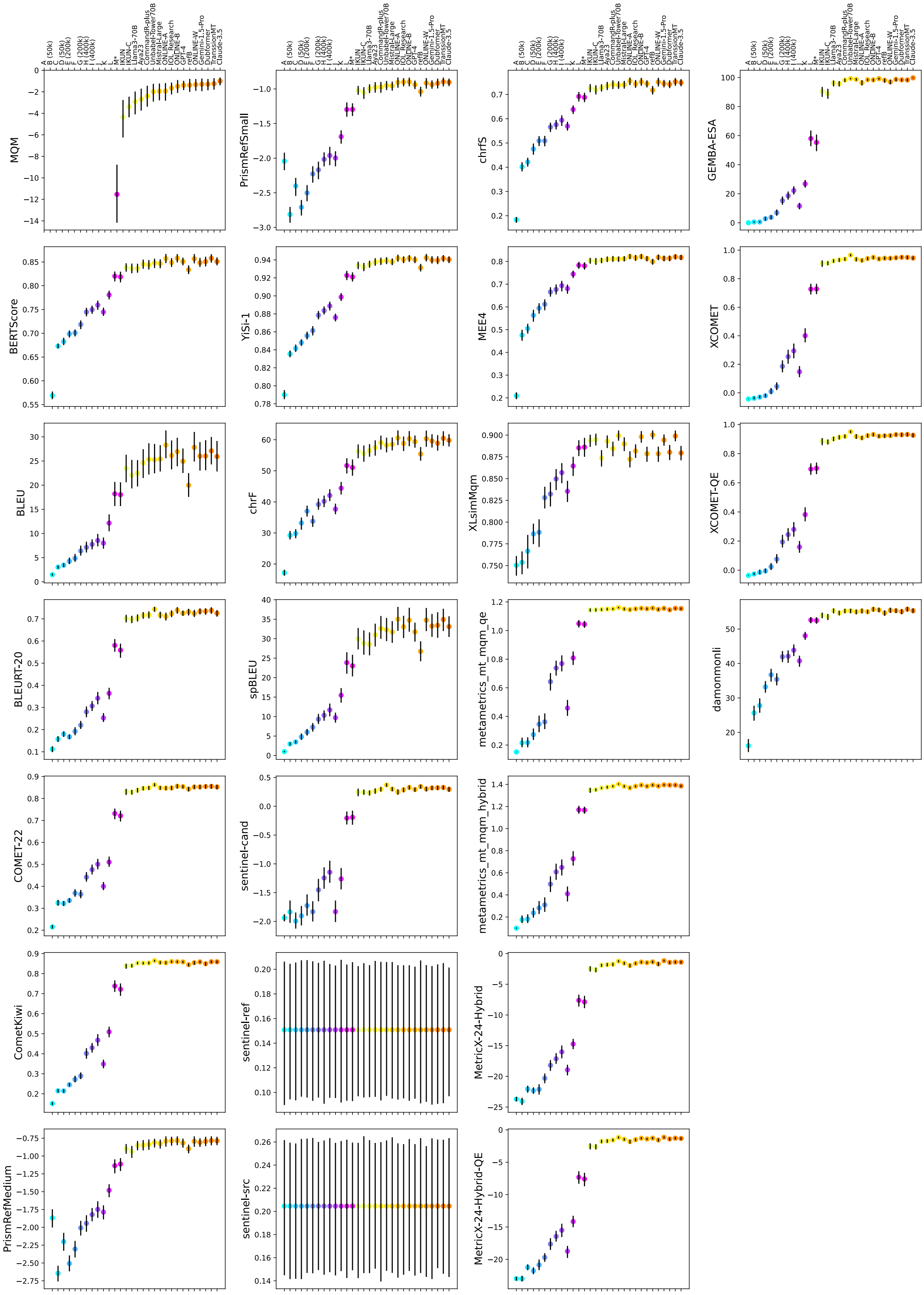

Figure 5: System average scores for English→German.

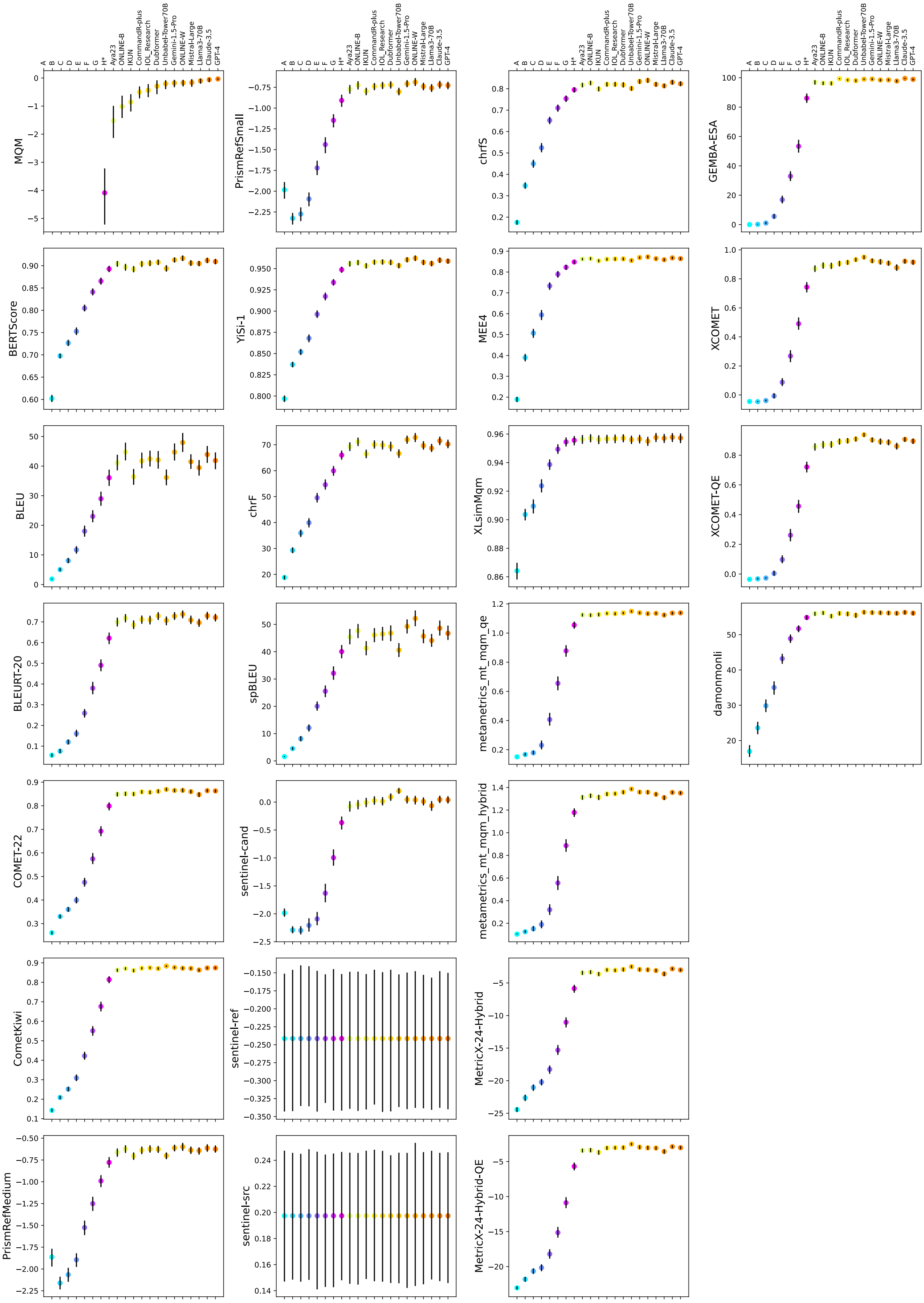

Figure 6: System average scores for English→Spanish.

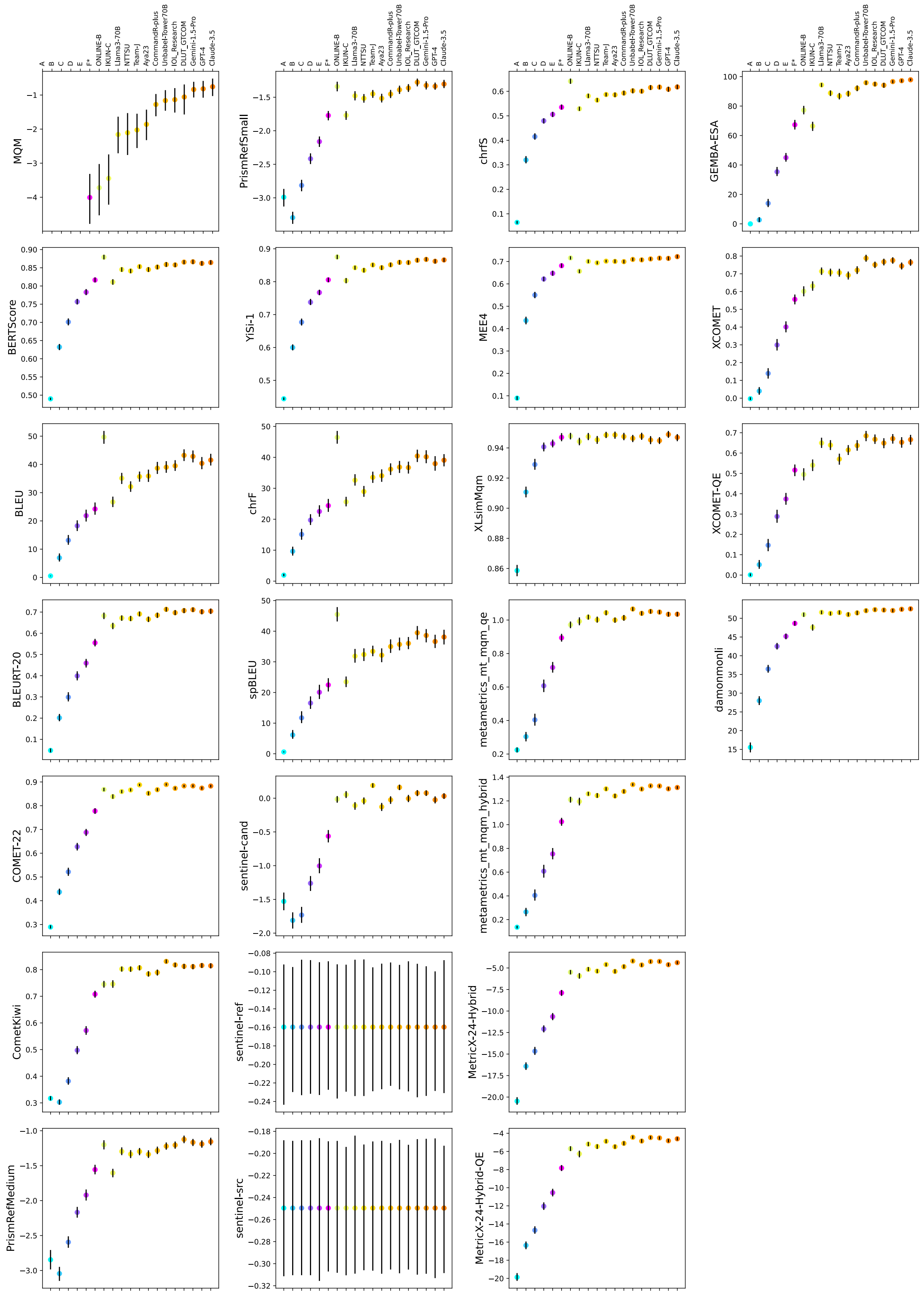

Figure 7: System average scores for Japanese→Chinese.

Metric	**Mix≥Wrong**
BERTScore	5
BLEU	5
BLEURT-20	5
COMET-22	5
CometKiwi	0
PrismRefMedium	0
PrismRefSmall	0
YiSi-1	5
chrF	5
spBLEU	5
chrfS	5
MEE4	5
XLsimMqm	5
MetaMetrics-MT-QE	0
MetaMetrics-MT	4
MetricX-24-Hybrid	1
MetricX-24-Hybrid-QE	0
GEMBA-ESA	0
XCOMET	0
XCOMET-QE	0
damonmonli	3

Table 9: `jpn`→`zho`: Number of times (out of 5) that the metric scored a mixed-language hypothesis higher than or equal its wrong-language pair.

D.2 Remaining Additional Plots

For other examples of of the empty string plots, as well as for additional plots showing histograms and scatterplots, see https://github.com/nrc-cnrc/MSLC.