Machine Translation Meta Evaluation through Translation Accuracy Challenge Sets

Nikita Moghe
School of Informatics, University of
Edinburgh*

Arnisa Fazla
Department of Computational Linguistics,
University of Zurich

Chantal Amrhein
Textshuttle, Zurich

Tom Kocmi
Microsoft

Mark Steedman
School of Informatics, University of
Edinburgh

Alexandra Birch
School of Informatics, University of
Edinburgh

Rico Sennrich
Department of Computational Linguistics,
University of Zurich

Liane Guillou
Department of Computer Science, RISE
Research Institutes of Sweden

Recent machine translation (MT) metrics calibrate their effectiveness by correlating with human judgement. However, these results are often obtained by averaging predictions across large test sets without any insights into the strengths and weaknesses of these metrics across different error types. Challenge sets are used to probe specific dimensions of metric behaviour but there are very few such datasets and they either focus on a limited number of phenomena or a limited number of language pairs. We introduce ACES, *a contrastive challenge set spanning 146 language pairs, aimed at discovering whether metrics can identify 68 translation accuracy errors. These phenomena range from basic alterations at the word/character level to more intricate errors based on discourse and real-world knowledge. We conducted a large-scale study by benchmarking* ACES *on 50 metrics submitted to the WMT 2022 and 2023 metrics shared tasks. We benchmark metric performance, assess their incremental performance over successive campaigns, and measure their sensitivity to a range of linguistic phenomena. We also investigate claims that Large Language Models (LLMs) are effective as MT evaluators, addressing the limitations of previous studies by providing a more holistic evaluation that covers a range of linguistic phenomena and language pairs and includes both low- and medium-resource languages. Our results demonstrate that different metric families struggle with different phenomena and that LLM-based methods fail to demonstrate reliable performance. Our analyses indicate that most metrics ignore the source sentence, tend to prefer surface-level overlap and end up incorporating properties of base models which are not always beneficial. To further encourage detailed evaluation beyond singular scores, we expand* ACES *to include error span annotations, denoted as* SPAN-ACES *and we use this dataset to evaluate span-based error metrics showing these metrics also need considerable improvement.*

∗ Corresponding author: nikita.moghe@ed.ac.uk

Finally, we provide a set of recommendations for building better MT metrics, including focusing on error labels instead of scores, ensembling, designing strategies to explicitly focus on the source sentence, focusing on semantic content rather than relying on the lexical overlap, and choosing the right base model for representations.

1. Introduction

Machine Translation (MT) metrics are a fundamental component of the development of high-quality MT systems as most state-of-the-art models claim their effectiveness through such metrics (Kocmi et al. 2021). While human evaluation of these MT systems is ideal, it is labour-intensive, time-consuming, and expensive. Development of automatic metrics has thus received significant interest over the past years (Koehn and Monz 2006; Freitag et al. 2023) resulting in a surge of new metrics. These metrics are typically judged by their ability to distinguish the quality of one machine translation system over another (system-level) on large test sets. This type of evaluation only provides an overview and it is difficult to identify whether these metrics are robust to specific MT errors.

To systematically study the advantages and shortcomings of MT metrics, and to identify broad trends in metric development, we rely on the construction of challenge sets for MT metrics. Challenge sets are a useful tool in measuring the performance of systems or metrics on one or more specific phenomena of interest. They may be used to compare the performance of a range of *different* systems or to identify performance improvement/degradation between successive iterations of the *same* system.

The WMT 2021 Metrics shared task (Freitag et al. 2021b) introduced a shared task on constructing contrastive challenge sets for the evaluation of MT metrics. Contrastive challenge sets aim to assess how well a given metric can discriminate between a *good* and *incorrect* translation of the *source* text where the incorrect translation consists of a translation error of interest. Providing a *reference* translation allows for flexibility: it may be included to assess reference-based (i.e. MT) metrics or excluded to assess reference-free (i.e. Quality Estimation (QE)) metrics. Benchmarking metrics on such challenge sets provides insights into their strengths while simultaneously uncovering their weaknesses on different translation errors.

In this work, we describe the Translation **A**ccuracy **C**halleng**E S**et (ACES) dataset submitted to the challenge sets subtask of the WMT 2022 and 2023 Metrics shared task and its subsequent expansion to include error span annotations (SPAN-ACES). The ACES dataset[1] (Amrhein, Moghe, and Guillou 2022) consists of 36,476 examples covering 146 language pairs and representing challenges from 68 phenomena. Most MT metric challenge sets (Avramidis et al. 2018; Alves et al. 2022; Karpinska et al. 2022) either focus on a small number of phenomena or a small number of languages. Our datasets are large scale in coverage of phenomena as well as language pairs providing a comprehensive challenge set for MT metrics.

We focus on translation accuracy errors because in recent years, machine translation outputs have become increasingly fluent (Bentivogli et al. 2016; Toral and Sánchez-Cartagena 2017; Castilho et al. 2017). Further, accuracy errors can have dangerous consequences in certain contexts, for example in the medical and legal domains (Vieira, O'Hagan, and O'Sullivan 2021).

ACES uses the hierarchy of errors under the class *Accuracy* from the Multidimensional Quality Metrics (MQM) ontology (Lommel, Burchardt, and Uszkoreit 2014) to design the ACES challenge sets. We extend this ontology by two error classes (translations defying real-world knowledge and translations in the wrong language) and specify several more specific subclasses such as discourse-level errors or ordering mismatches. We include phenomena ranging from

---

1 The ACES dataset is available at `https://huggingface.co/datasets/nikitam/ACES`
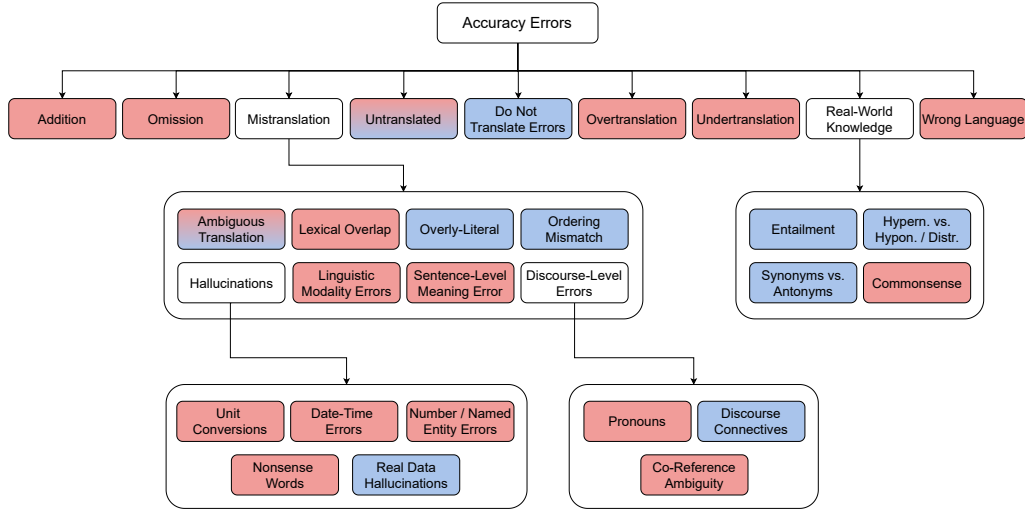
Figure 1
Diagram of the error categories on which our collection of challenge sets is based. Red means challenge sets are created automatically, blue means challenge sets are created manually.

simple perturbations involving the omission/addition of characters or tokens to more complex examples involving mistranslation e.g. ambiguity and hallucinations in translation, untranslated elements of a sentence, discourse-level phenomena, and real-world knowledge. A full overview of all error classes can be seen in Figure 1. Our challenge set consists of synthetically generated adversarial examples, examples from re-purposed contrastive MT test sets (both marked in red), and manually annotated examples (marked in blue).

We use ACES to benchmark the metrics that participated in the WMT 2022 and 2023 metrics shared tasks. We also investigate whether Large Language Models can perform MT evaluation (Kocmi and Federmann 2023b; Xu et al. 2023). We conduct several analyses on these results revealing:

1. There is no *winning* metric as conducting granular evaluation reveals different metrics have different strengths and weaknesses.

2. Most metrics tend to disregard information present in the source.

3. Reference-based neural metrics still rely on surface-level overlap.

4. Some properties of the base model in neural metrics may cause undesirable effects on evaluation.

A metric that can, in addition to providing scores, accurately label errors in MT output provides many clear advantages over one that only provides scores (Freitag et al. 2021a). Observations by Moghe et al. (2023) suggest that interpreting the quality of MT output based on scores is both unreliable and uninformative. Instead, they recommend the development of metrics that predict labels for error spans in the MT output. Similarly, Lommel, Burchardt, and Uszkoreit (2014); Freitag et al. (2021a) and the recent WMT challenges (Freitag et al. 2021b, 2022, 2023) also advocate the use of labelled error spans for MT evaluation. When considering whether to deploy an MT system (or which of several systems to deploy), system developers can take into consideration the type, frequency, and severity of the errors that the system is likely to make,

coupled with information about what types of errors may be tolerated/not for a given downstream task.

With these motivations, we extend the ACES dataset into SPAN-ACES, where we include error span annotations for each example. These annotations indicate the location of error spans present in the *incorrect* translation and pertaining to the specific linguistic phenomenon in focus. Whilst some currently available MT metrics are already able to mark error spans including MATESE (Perrella et al. 2022a), COMET-22 (Rei et al. 2022) trained on MQM (Lommel, Burchardt, and Uszkoreit 2014), GEMBA-MQM (Kocmi and Federmann 2023a), AutoMQM (Fernandes et al. 2023) that prompt LLMs to obtain the corresponding error span, we believe that error-span labelling is an important next step in MT metric evolution. Independent challenge sets such as SPAN-ACES will be essential in driving development forward. We benchmark GEMBA-MQM (Kocmi and Federmann 2023a)), XCOMET-XL (Guerreiro et al. 2023), and adapted versions of COMET-22 (Rei et al. 2022) and UniTE (Wan et al. 2022b) on SPAN-ACES.

In this article, we provide an overview of the ACES challenge set and its participation at the WMT 2022 and 2023 Metrics shared task - Challenge Sets subtask (Amrhein, Moghe, and Guillou 2022, 2023). We list our contributions and also report on extensions to our previously published work, including novel contributions (listed from 4):

- We briefly present the construction of ACES, containing 36k examples across 146 language pairs and 68 phenomena.

- We evaluate ACES on the metrics submitted to the WMT 2022 and WMT 23 Metrics shared task providing an overview of the performance of 50 different metrics.

- We conduct several analyses on these metrics revealing their drawbacks and also providing recommendations to mitigate them.

- We describe the construction of SPAN-ACES, an extended version ACES which includes error span annotations.

- Using SPAN-ACES, we benchmark the performance of currently available metrics for the task of labelling errors in MT output. Our results suggest that these methods show some success on the error labelling task with the highest span-F1 score reaching 26.9. However, these results and corresponding poor results on the contrastive task also raise new questions in labelling MT errors as evaluation.

- We present the results of analyses aimed at determining how *sensitive* metrics are to different phenomena. This is grounded in our assertion that an ideal metric should assign comparatively high scores to good translations and low scores to incorrect translations.

- We investigate claims that Large Language Models (LLMs) may be used as MT evaluators and describe experiments on LLMs from three different families. Benchmarking these LLMs on ACES reveals that these models perform worse than the string-overlap metrics. These results degrade further in the reference-free setting where all of the LLMs have a negative correlation across all of the ACES categories.

We advocate steering metric development towards methods that produce error labels in addition to the scores. Based on our analyses, we also recommend that metric developers consider: a) combining metrics with different strengths, e.g. in the form of ensemble models, b) paying

more attention to the source and avoiding over-reliance on surface-overlap with the reference, and c) check the base model properties prior to their use in developing new metrics

We propose the adoption of both ACES and SPAN-ACES by the MT community, as a benchmark for developing Machine Translation metrics. We envisage several use cases in which the challenge sets may be employed: to profile and compare metric performance across a range of error categories, and to identify improvement/degradation in performance of successive development iterations of the same metric. Similarly, MT models can also be evaluated using this dataset by calculating sentence-level perplexity of the two translations. Furthermore, we propose the use of SPAN-ACES to aid in advancing the development of the next generation of MT metrics which aim to provide error-span labels over MT output in addition to scores.

2. Related Work

Challenge sets have been used for a range of NLP tasks to investigate the behaviour of these tasks under a specific phenomenon rather than the standard test distribution (Popović and Castilho 2019). Challenge sets aim to provide insights on whether state-of-the-art models are robust to domain shifts, simple textual perturbations, whether they have some understanding of linguistic phenomena such as negation/commonsense or simply rely on shallow heuristics, to name a few. The earliest introduction of challenge sets was by King and Falkedal (1990) who probed the acceptability of machine translations for different domains. Since then challenge sets have been developed for different fields within NLP including parsing (Rimell, Clark, and Steedman 2009), NLI (McCoy and Linzen 2019; Rocchietti et al. 2021), question answering (Ravichander et al. 2021), reading comprehension (Khashabi et al. 2018) and sentiment analysis (Li, Cohn, and Baldwin 2017; Mahler et al. 2017; Staliūnaitė and Bonfil 2017). Challenge sets are also referred as "adversarial datasets" which also create examples by perturbing the standard test set to fool the model (Smith (2012); Jia and Liang (2017), *inter-alia*).

Challenge sets for evaluating MT systems have focused on the translation models' ability to generate the correct translation given a phenomenon of interest. These include word sense ambiguity (Rios, Müller, and Sennrich 2018; Campolungo et al. 2022), gender bias (Rudinger, May, and Van Durme 2017; Zhao et al. 2018; Stanovsky, Smith, and Zettlemoyer 2019), structural divergence (Isabelle, Cherry, and Foster 2017) and discourse level phenomena (Guillou and Hardmeier 2016; Emelin and Sennrich 2021). While such challenge sets focus on evaluating specific MT models, it is necessary to identify whether the existing MT evaluation metrics also perform well under these and related phenomena. Following the success of neural MT metrics, which have been shown to correlate well with human judgements (Freitag et al. 2021b; Kocmi et al. 2021), the development of challenge sets designed to examine their strengths and weaknesses has received considerable interest. However, metric weaknesses remain relatively unknown and only a small number of works (e.g. Hanna and Bojar (2021) and Amrhein and Sennrich (2022)) have proposed systematic analyses to uncover them.

Early work on constructing challenge sets for metric evaluation typically focused on a small range of phenomena (Specia et al. 2020), synthetic perturbations (Freitag et al. 2021b), or manual perturbations for high-resource language pairs (Avramidis et al. 2018). These limitations have been addressed in the development of the DEMETR (Karpinska et al. 2022) and ACES datasets.

DEMETR (Karpinska et al. 2022), which comprises 31K English examples translated from ten languages, was developed for evaluating MT metric sensitivity to a range of 35 different types of linguistic perturbations, belonging to semantic, syntactic, and morphological error categories. These were divided into minor, major, and critical errors according to the type of perturbation, similar to the grading of error categories to compute the weighted ACES-Score. As in ACES, example generation was carefully designed to form minimal pairs such that the perturbed translation only differs from the actual translation in one aspect. The application of

DEMETR in evaluating a suite of baseline metrics revealed a similar pattern to the analyses in Amrhein, Moghe, and Guillou (2022) - that metric performance varies considerably across the different error categories, often with no clear winner. It is worth noting that DEMETR and ACES each have their respective advantages: all examples in DEMETR have been verified by human annotators; ACES provides broader coverage in terms of both languages and linguistic phenomena.

In addition to ACES, three other datasets were submitted to the WMT 2022 challenge sets shared task (Freitag et al. 2022): SMAUG (Alves et al. 2022), the HWTSC challenge set (Chen et al. 2022), and the DFKI challenge set (Avramidis and Macketanz 2022). These datasets differ from ACES in terms of their size, and the languages and phenomena/categories they cover. Both SMAUG and HWTSC are relatively small datasets (<1000 examples) focusing on a small set of five phenomena, each pertaining to a single category of critical error for meaning change. In comparison, the DFKI challenge set is much larger – it contains 19,347 examples and covers over 100 linguistically motivated phenomena, which are organised into 14 categories. Whereas the aim of ACES was to provide a broad coverage of language pairs, the other datasets provide an in-depth focus on specific high-resource language pairs: SMAUG (pt↔en and es→en), DFKI (de↔en), and HWTSC (zh↔en). Whilst there is a clear overlap between the ACES phenomena and those in SMAUG and HWTSC, many of the phenomena in the DFKI dataset are complementary such that in the case of evaluating metrics for the German-English pair, metric developers might consider benchmarking on both datasets.

The WMT 2023 Challenge Sets submissions included ACES, MSLC23 (Lo, Larkin, and Knowles 2023), and an extended version of the DFKI challenge set to include the en→ru language pair plus additional examples and phenomena for the en→de language pair (Avramidis et al. 2023). The MSLC23 dataset covers four language pairs (zh→en, he↔en, and en→de) and includes examples of low-, medium- and high-quality output designed to provide an interpretation of metric performance across a range of different levels of translation quality. The motivation for this is that whilst metric performance may be evaluated on high-quality MT output, these same metrics may later be used to evaluate low-quality MT output, and it is, therefore, important to understand their performance in the lower-quality setting.

We note that our contrastive challenge set with error span annotations is the first of its kind.

3. Challenge Sets

Creating a contrastive challenge set for evaluating a machine translation evaluation metric requires a source sentence, a reference translation, and two translation hypotheses: one which contains an error or phenomenon of interest (the "incorrect" translation) and one which is a correct translation in that respect (the "good" translation). One possible way to create such challenge sets is to start with two alternative references (or two identical copies of the same reference) and insert errors into one of them to form an incorrect translation while the uncorrupted version can be used as the good translation. This limits the full evaluation scope to translation hypotheses that only contain a single error. To create a more realistic setup, we also create many challenge sets where the good translation is not free of errors, but it is a better translation than the incorrect translation. For automatically created challenge sets, we put measures in place to ensure that the incorrect translation is indeed a worse translation than the good translation.

3.1 Datasets

The examples in ACES are based on several academic datasets designed to test particular properties in Machine Translation or other multilingual NLP tasks. The majority of the examples in our challenge set were based on data extracted from three main datasets: FLORES-101, PAWS-

X, and XNLI (with additional translations from XTREME). **FLORES-101** (Goyal et al. 2022) and **FLORES-200** (NLLB Team et al. 2022) are low resource MT evaluation benchmarks with parallel data in 101 and 200 languages respectively. **PAWS-X** (Yang et al. 2019) is a cross-lingual dataset for paraphrase identification in seven languages that consists of pairs of sentences that are labelled as true or adversarial paraphrases. **XNLI** (Conneau et al. 2018) is a multilingual Natural Language Inference (NLI) dataset consisting of premise-hypothesis pairs with their corresponding inference label for 14 languages. The other datasets used in the development of ACES serve specific challenges. **WinoMT** (Stanovsky, Smith, and Zettlemoyer 2019), a challenge set developed for analysing gender bias in MT with examples exhibiting an equal balance of male and female genders, and of stereotypical and non-stereotypical gender-role assignments (e.g., a female nurse vs. a female doctor). **MuCoW** (Raganato, Scherrer, and Tiedemann 2019) is a multilingual contrastive word sense disambiguation test suite for machine translation. The **WMT 2018 English-German pronoun translation evaluation test suite** (Guillou et al. 2018) contains examples of the ambiguous English pronouns *it* and *they* extracted from the TED talks portion of ParCorFull (Lapshinova-Koltunski, Hardmeier, and Krielke 2018). The **Europarl ConcoDisco** corpus (Laali and Kosseim 2017) comprises the English-French parallel texts from Europarl (Koehn 2005) over which automatic methods were used to perform discourse connective annotation of their sense types. **Wino-X** (Emelin and Sennrich 2021) is a parallel dataset of German, French, and Russian Winograd schemas, aligned with their English counterparts used to test commonsense reasoning and coreference resolution of MT models.

We will now discuss the different categories of challenge sets. We list some examples from ACES in Table 1.

3.2 Addition and Omission

We create a challenge set for addition and omission errors which are defined in the MQM ontology as "target content that includes content not present in the source" and "errors where content is missing from the translation that is present in the source", respectively. We focus on the level of constituents and use an implementation by Vamvas and Sennrich (2022) to create synthetic examples of addition and omission errors using the likelihood of tokens for a given MT model. To generate examples, we use the concatenated dev and devtest sets from the FLORES-101 evaluation benchmark for 46 languages. We focus on the 46 languages for which there exists a stanza parser[2] and create datasets for all languages paired with English plus ten additional language pairs that we selected randomly. For translation, we use the M2M100[3] model with 1.2B parameters (Fan et al. 2021).

3.3 Mistranslation

The mistranslation phenomenon is broadly defined as the target translation not accurately containing the information in the source content.

3.3.1 Mistranslation - Ambiguous Translation. This error type is defined in the MQM ontology as a case where "an unambiguous source text is translated ambiguously". For this error type, we create challenge sets where MT metrics are presented with an unambiguous source and an ambiguous reference. The metrics then need to choose between two disambiguated translation hypotheses where only one meaning matches the source sentence. Therefore, these challenge sets

---

2 https://stanfordnlp.github.io/stanza/available_models.html
3 https://huggingface.co/facebook/m2m100_1.2B

test whether metrics consider the source when the reference is not expressive enough to identify the better translation. Since many reference-based metrics, by design, do not include the source to compute evaluation scores, we believe that this presents a challenging test set.

Our method for creating examples is inspired by Vamvas and Sennrich (2021) who score a translation against two versions of the source sentence, one with an added correct disambiguation cue and one with a wrong disambiguation cue to determine whether a translation model produced the correct translation or not. Instead of adding the disambiguation cues to the source, we use an unambiguous source and add disambiguation cues to an ambiguous reference to create two contrasting translation hypotheses. We create three separate challenge sets of this type:

Occupation Name Gender using the WinoMT dataset where the source language is a gendered language and the target language is English. The cues added to the reference to form the "good" and "incorrect" translations are "female" and "male".

Word Sense Disambiguation using the MuCoW dataset where the ambiguity lies in homographs in the target language that are unambiguous in the source sentence. The cues added to the reference to form the contrastive translations are sense-specific.

Discourse Connectives using the Europarl ConDisco corpus where the ambiguity lies in the English discourse connective "since" which can have both causal and temporal meanings.

3.3.2 Mistranslation - Hallucinations. In this category, we group several subcategories of mistranslation errors that happen at the word level and could occur due to hallucination by an MT model. Hallucinations are erroneous generations where the output is partially related or unrelated to the source sentence (Dale et al. 2023). These challenge sets test whether the machine translation evaluation metrics can reliably identify hallucinations when presented with a correct alternative translation.

We create five different challenge sets based on hallucination errors:

Date-Time Errors: using the FLORES-101 data where a month name in the reference (e.g. November) is replaced with a corresponding abbreviation in the "good" translation (e.g. Nov.) and a different month name in the "incorrect" translation (e.g. August).

Numbers and Named Entities: We create a challenge set for numbers and named entities where we perform character-level edits (adding, removing or substituting digits in numbers or characters in named entities) as well as word-level edits (substituting whole numbers or named entities). In the 2021 WMT metrics shared task, number differences were not a big issue for most neural metrics (Freitag et al. 2021b). However, we believe that simply changing a number in an alternative translation and using this as an incorrect translation as done by Freitag et al. (2021b) is an overly simplistic setup and does not cover the whole translation hypothesis space. To address this shortcoming, we propose a three-level evaluation. The first, easiest level follows Freitag et al. (2021b) and applies a change to an alternative translation to form an incorrect translation. The second level uses an alternative translation that is lexically very similar to the reference as the good translation and applies a change to the reference to form an incorrect translation. The third, and hardest level, uses an alternative translation that is lexically very different from the reference as the good translation and applies a change to the reference to form an incorrect translation. In this way, our challenge set tests whether the number and named entity differences can still be detected as the surface similarity between the two translation candidates decreases and the surface similarity between the incorrect translation and the reference increases. We use cross-lingual paraphrases from the PAWS-X dataset as a pool of alternative translations to create this challenge set. We only consider language pairs for which we can use a spacy NER model on the target side, which results in 42 language pairs.

Unit Conversion: using FLORES-101 dataset, where we replace unit mentions in the reference (e.g. 100 feet) with a different unit and corresponding amount in the "good" translation (e.g. 30.5

metres) and either the wrong amount (e.g. 100 metres) or wrong unit (30.5 feet) compared to the reference in the "incorrect" translation.

Nonsense Words: We develop a challenge set for evaluating hallucinations at subword level (Sennrich, Haddow, and Birch 2016). To create this challenge set, we consider tokens which are broken down into at least two subwords and then randomly swap those subwords with other subwords to create nonsense words by using the multilingual BERT tokenizer (Devlin et al. 2019). We use the paraphrases from the PAWS-X dataset as good translations and randomly swap one subword in the reference to generate an incorrect translation.

Real Data Hallucinations: To also create a more realistic hallucination benchmark, we manually check some machine translations of the FLORES-101 dev and devtest sets for four language pairs: de→en, en→de, fr→de and en→mr. We consider both cases where a more frequent, completely wrong word occurs and cases where the MT model started with the correct subword but then produced random subwords as hallucinations. Translations with a hallucination are used as incorrect translations. We manually replace the hallucination part with its correct translation to form the good translation.

3.3.3 Mistranslation - Lexical Overlap.
Language models trained with the masked language modelling objective are successful on downstream tasks because they model higher-order word co-occurrence statistics instead of syntactic structures (Sinha et al. 2021). Similarly, existing surface-level metrics rely on n-gram matching between the hypothesis and the reference. We create this challenge set to test if metrics can reliably identify an incorrect translation especially when it shares a high degree of lexical overlap with the reference. To create such examples, we use the PAWS-X dataset for which adversarial paraphrase examples were constructed by changing the word order and/or the syntactic structure while maintaining a high degree of lexical overlap.

3.3.4 Mistranslation - Linguistic Modality.
Modal auxiliary verbs signal the function of the main verb that they govern. For example, they may be used to denote possibility ("could"), permission ("may"), the giving of advice ("should"), or necessity ("must"). We are interested in whether MT evaluation metrics can identify when modal auxiliary verbs are incorrectly translated. We focus on the English modal auxiliary verbs: "must" (necessity), and "may", "might", "could" (possibility). We then translate the source sentence using Google Translate to obtain the "good" translation and manually replace the modal verb with an alternative with the same meaning where necessary (e.g. "have to" denotes necessity as does "must"; also "might", "may" and "could" are considered equivalent). For the incorrect translation, we manually substitute the modal verb that conveys a different meaning or *epistemic strength* e.g. in the example above "might" (possibility) is replaced with "will", which denotes (near) certainty. We use a combination of the FLORES-200 and PAWS-X datasets as the basis of the challenge sets.

3.3.5 Mistranslation - Overly Literal Translations.
MQM defines this error type as translations that are overly literal, for example, literal translations of figurative language. We create two challenge sets based on this error type:

Idioms: We create this challenge set based on the PIE[4] parallel corpus of English idiomatic expressions and literal paraphrases (Zhou, Gong, and Bhat 2021). We manually translate 102 parallel sentences into German for which we find a matching idiom that is not a word-by-word translation of the original English idiom. Further, we create an overly literal translation of the English and German idioms. We use either the German or English original idiom as the source sentence. Then, we either use the correct idiom in the other language as the reference and the

---

4 https://github.com/zhjjn/MWE_PIE

literal paraphrase as the good translation, or vice versa. The incorrect translation is always the overly literal translation of the source idiom.

Real Data Errors: For this challenge set, we manually check MT translations of the FLORES-101 datasets. If we find an overly-literal translation, we manually correct it to form the good translation and use the overly-literal translation as the incorrect translation.

3.3.6 Mistranslation - Sentence-Level Meaning Error. We also consider a special case of sentence-level semantic error that arises due to the nature of the task of Natural Language Inference (NLI). The task of NLI requires identifying where the given hypothesis is an entailment, contradiction, or neutral, for a given premise. Thus, the premise and hypothesis have substantial overlap but they vary in meaning. We use the XNLI dataset to create such examples where there is at least a 0.5 chrF score between the English premise and hypothesis only for the neutral and contradiction examples. We use either the premise/hypothesis as the reference, an automatic translation as the "good translation", premise/hypothesis from the remaining non-English languages, and hypothesis/premise as the "incorrect translation".

3.3.7 Mistranslation - Ordering Mismatch. We also investigate the effects of changing word order in a way that changes meaning. For example, "I like apple pie and fried chicken" is changed to " I like chicken pie and fried apple" to form the incorrect translation. This challenge set is created manually by changing translations from the FLORES-101 dataset and covers de→en, en→de and fr→de.

3.4 Mistranslation - Discourse-level Errors

We introduce a new subclass of mistranslation errors that specifically cover discourse-level phenomena. We create several challenge sets based on discourse-level errors:

Pronouns: To create these challenge sets, we use the English-German pronoun translation evaluation test suite from the WMT 2018 shared task as the basis for our examples. We focus on the following six categories derived from the manually annotated pronoun function and attribute labels: pleonastic *it*, anaphoric subject and non-subject position *it*, anaphoric *they*, singular *they*, and group *it/they*. We use the MT translations as the "good" translations and automatically generate "incorrect" translations using one of the following strategies: *omission* - the translated pronoun is deleted from the MT output, *substitution* - the "correct" pronoun is replaced with an "incorrect" form.

Discourse Connectives: We leverage the Europarl ConcoDisco corpus of parallel English/French sentences with discourse connectives marked and annotated for sense, and select examples with ambiguity in the French source sentence. We construct the good translation by replacing instances of "while" (temporal) with "as" or "as long as" and instances of "while" (comparison) as "whereas" (ensuring grammaticality is preserved). For the incorrect translation, we replace the discourse connective with one with the alternative sense of "while" e.g. we use "whereas" (comparison) where a temporal sense is required.

Commonsense Co-Reference Disambiguation: We use the English sentences in the Wino-X challenge set which were sampled from the Winograd schema. All contain the pronoun *it* and were manually translated into two contrastive translations for de, fr, and ru. Based on this data, we create our challenge sets covering two types of examples: For the first, the good translation contains the pronoun referring to the correct antecedent, while the incorrect translation contains the pronoun referring to the incorrect antecedent. For the second, the correct translation translates the instance of *it* into the correct disambiguating filler, while the second translation contains the pronoun referring to the incorrect antecedent.

10

3.5 Untranslated

MQM defines this error type as "errors occurring when a text segment that was intended for translation is left untranslated in the target content". We create two challenge sets based on untranslated content errors:

Word-Level: We manually annotate real errors in translations of the FLORES-101 dev and devtest sets. We count complete copies as untranslated content as well as content that comes from the source language but was only adapted to look more like the target language.

Sentence-Level: We create a challenge set for untranslated sentences by simply copying the entire source sentence as the incorrect translation. We used a combination of examples from the FLORES-200, XNLI, and PAWS-X datasets to create these examples.

3.6 Do Not Translate Errors

This category of errors is defined in MQM as content in the source that should be copied to the output in the source language but was mistakenly translated into the target language. Common examples of this error type are company names or slogans. Here, we manually create a challenge set based on the PAWS-X data which contains many song titles that should not be translated. To construct the challenge set, we use one paraphrase as the good translation and manually translate an English sequence of tokens (e.g. a song title) into German to form the incorrect translation.

3.7 Overtranslation and Undertranslation

Hallucinations from a translation model can often produce a term which is either more generic than the source word or more specific. Within the MQM ontology, the former is referred to as undertranslation while the latter is referred to as overtranslation. For example, "car" may be substituted with "vehicle" (undertranslation) or "BMW" (overtranslation). A randomly selected noun from the reference translation is replaced by its corresponding hypernym or hyponym (by using Wordnet) to simulate undertranslation or overtranslation errors, respectively.

3.8 Real-world Knowledge

We propose a new error category where translations disagree with real-world knowledge in addition to the accuracy categories in MQM. We create five challenge sets based on this error type. For the first four, we manually construct examples each for en→de and de→en. We used German-English examples from XNLI, plus English translations from XTREME as the basis for our examples. Typically, we select a single sentence, either the premise or hypothesis from XNLI, and manipulate the MT translations.

Textual Entailment: We construct examples for which the good translation entails the meaning of the original sentence (and its reference). For example, we use the entailment *was murdered →
died* (i.e. if a person is murdered then they must have died) to construct the good translation in the example above. We construct the incorrect translation by replacing the entailed predicate (*died*) with a related but non-entailed predicate (here *was attacked*) – a person may have been murdered without being attacked, i.e. by being poisoned for example.

Hypernyms and Hyponyms: We consider a translation that contains a *hypernym* of a word to be better than one that contains a *hyponym*. For example, whilst translating "Hund" ("dog") with the broader term "animal" results in some loss of information, this is preferable over hallucinating information by using a more specific term such as "labrador" (i.e. an instance of the hyponym

class "dog"). We used Wordnet and WordRel.com[5] (an online dictionary of words' relations) to identify hypernyms and hyponyms of nouns within the reference sentences, and used these as substitutions in the MT output: hypernyms are used in the "good" translations and hyponyms in the "incorrect" translations.

Hypernyms and Distractors: Similar to above, we construct examples in which the good translation contains a hypernym (e.g. "pet") of the word in the reference (e.g. "dog"). We form the incorrect translation by replacing the original word in the source/reference with a different member from the same class (e.g. "cat"; both cats and dogs belong to the class of pets). Note the techniques in Section 3.7 manipulate the reference only to create an incorrect translation with the respective error.

Antonyms: We also construct incorrect translations by replacing words with their corresponding antonyms from Wordnet. We construct challenge sets for both nouns and verbs. For nouns, we automatically constructed incorrect translations by replacing nouns in the reference with their antonyms. In the case of verbs, we manually constructed a more challenging set of examples intended to be used to assess whether the metrics can distinguish between translations that contain a synonym versus an antonym of a given word.

Commonsense: We are also interested in whether evaluation metrics prefer translations that adhere to common sense. To test this, we remove explanatory subordinate clauses from the sources and references in the dataset described in Section 3.4. This guarantees that when choosing between a good and incorrect translation, the metric cannot infer the correct answer from looking at the source or the reference. We then pair the shortened source and reference sentences with the full translation that follows commonsense as the good translation and the full translation with the other noun as the incorrect translation.

3.9 Wrong Language

Most of the representations obtained from large multilingual language models do not explicitly use the language identifier (id) as an input while encoding a sentence. Here, we are interested in checking whether sentences which have similar meanings are closer together in the representation space of neural MT evaluation metrics, irrespective of their language. We create a challenge set for embedding-based metrics using the FLORES-200 dataset where the incorrect translation is in a similar language (same typology/same script) to the reference (e.g. a Catalan translation may be used as the incorrect translation if the target language is Spanish).

3.10 Fluency

Although the focus of ACES is on accuracy errors, we also include a small set of fluency errors for the punctuation category.

Punctuation: We assess the effect of deleting and substituting punctuation characters. We employ four strategies: 1) deleting all punctuation, 2) deleting only quotation marks (i.e. removing indications of quoted speech), 3) deleting only commas (i.e. removing clause boundary markers), 4) replacing exclamation points with question marks (i.e. statement → question). In strategies 1 and, especially, 3 and 4, some of the examples may also contain accuracy-related errors. For example, the meaning of the sentence could be changed in the incorrect translation if we remove a comma, e.g. in the (in)famous example "Let's eat, Grandma!" vs. "Let's eat Grandma!". We use the TED Talks from the WMT 2018 English-German pronoun translation evaluation test suite and apply all deletions and substitutions automatically.

	Addition
	target includes content not present in the source
SRC (de):	In den letzten 20 Jahren ist die Auswahl in Uptown Charlotte exponentiell gewachsen.
REF (en):	In the past 20 years, the amount in Uptown Charlotte has grown exponentially.
✓:	Over the past 20 years, the selection in Uptown Charlotte has grown exponentially.
✗:	Over the past 20 years, the selection of **child-friendly options** in Uptown Charlotte has grown exponentially.
	Omission
	errors where content is missing from the translation that is present in the source
SRC (fr):	Une tornade est un tourbillon d'air à basse-pression en forme de colonne, l'air alentour est aspiré vers l'intérieur et le haut.
REF (en):	A tornado is a **spinning column** of very low-pressure air, which sucks the surrounding air inward and upward.
✓:	A tornado is a **column-shaped** low-pressure air turbine, the air around it is sucked inside and up.
✗:	A tornado is a low-pressure air turbine, the air around it is sucked inside and up.
	Untranslated - Word Level
	errors occurring when a text segment that was intended for translation is left untranslated in the target content
SRC (fr):	À l'origine, l'émission mettait en scène des **comédiens de doublage** amateurs, originaires de l'est du Texas.
REF (de):	Die Sendung hatte ursprünglich lokale Amateur**synchronsprecher** aus Ost-Texas.
✓ (copy):	Ursprünglich spielte die Show mit Amateur**synchronsprechern** aus dem Osten von Texas.
✓ (syn.):	Ursprünglich spielte die Show mit Amateur-**Synchron-Schauspielern** aus dem Osten von Texas.
✗:	Ursprünglich spielte die Show mit Amateur-**Doubling-Schauspielern** aus dem Osten von Texas.
	Mistranslation - Ambiguous Translation
	an unambiguous source text is translated ambiguously
SRC (de):	Der Manager feuerte **die** Bäcke**rin**.
REF (en):	The manager fired the baker.
✓:	The manager fired the **female** baker.
✗:	The manager fired the **male** baker.
	Do Not Translate
	content in the source that should be copied to the output in the source language, but was mistakenly translated into the target language.
SRC (en):	Dance was one of the inspirations for the exodus - song **"The Toxic Waltz"**, from their 1989 album "Fabulous Disaster".
REF (de):	Dance war eine der Inspirationen für das Exodus-Lied **„The Toxic Waltz"** von ihrem 1989er Album „Fabulous Disaster".
✓:	Der Tanz war eine der Inspirationen für den Exodus-Song **„The Toxic Waltz"**, von ihrem 1989er Album „Fabulous Disaster".
✗:	Der Tanz war eine der Inspirationen für den Exodus-Song **„Der Toxische Walzer"**, von ihrem 1989er Album „Fabulous Disaster".
	Undertranslation
	erroneous translation has a meaning that is more generic than the source
SRC (de):	Bob und Ted waren Brüder. Ted ist der **Sohn** von John.
REF (en):	Bob and Ted were brothers. Ted is John's **son**.
✓:	Bob and Ted were brothers, and Ted is John's **son**.
✗:	Bob and Ted were brothers. Ted is John's **male offspring**.
	Overtranslation
	erroneous translation has a meaning that is more specific than the source
SRC (ja):	その 40 分の**映画**はアノーがアラン・ゴダードと協力して脚本を書いた。
REF (en):	The 40-minute **film** was written by Annaud with Alain Godard.
✓:	The 40-minute **film** was written by Annaud along with Alain Godard.
✗:	The 40-minute **cinema verite** was written by Annaud with Alain Godard.
	Real-world Knowledge - Textual Entailment
	meaning of the source/reference is entailed by the "good" translation
SRC (de):	Ein Mann **wurde ermordet**.
REF (en):	A man **was murdered**.
✓:	A man **died**.
✗:	A man **was attacked**.
	Wrong Language
	incorrect translation is a perfect translation in a related language
SRC (en):	Cell comes from the Latin word cella which means small room.
REF (es):	El término célula deriva de la palabra latina cella, que quiere decir «cuarto pequeño».
✓ (es):	La célula viene de la palabra latina cella que significa habitación pequeña.
✗ (ca):	Cèl·lula ve de la paraula llatina cella, que vol dir habitació petita.

Table 1

Examples from each top-level accuracy error category in ACES. An example consists of a source sentence (SRC), reference (REF), good (✓) and incorrect (✗) translations, language pair, and a phenomenon label. We also provide a description of the relevant phenomenon. en: English, de: German, fr: French, ja: Japanese, es: Spanish, ca: Catalan

See Appendix Appendix C: Distribution of Examples Across Language Pairs and Appendix D: Distribution of Language Pairs Across Phenomena for further information on the distribution of examples and language pairs in ACES.

4. Span Annotations

To support the development of Quality Estimation and MT evaluation metrics that predict error spans, we extended the original version of ACES (released at WMT 2022) to include error span annotations. Specifically, we annotated all error spans of the type denoted by the phenomenon category label, ignoring the presence of errors belonging to other categories. We therefore label only errors present in the incorrect translation, which by design contains errors of the phenomenon category denoted by the label. We annotate spans at the word/token level similar to the MQM format (Freitag et al. 2021a) and in line with recent developments in error span prediction metrics (Perrella et al. 2022a; Rei et al. 2022). Following the WMT 2022 MQM Human Evaluation span annotation format (Freitag et al. 2022), error spans are enclosed in tags (<v> error span </ v>) denoting the start and end position of the error in the incorrect translation. Note that due to the formulation of the manual annotation guidelines (see Section 9) it is not possible for two spans to overlap.

We provide annotations for all ACES examples, using a combination of automated and manual methods. The annotation methods used for each phenomenon can be found in Appendix E: ACES Annotation Methods per Phenomena. For many of the phenomena categories, we were able to automatically annotate examples using rule-based methods informed by the methodology that we followed to construct the examples. For the remaining phenomena, which we could not annotate automatically due to the manual methods used to generate the good and incorrect translations, we annotated the error spans manually (see Section 9). We also manually annotated a small number of examples (1959 from the mistranslation phenomena and 3 from the real-world knowledge phenomena) for which the automated annotation rules failed.

4.1 Automatic Annotations

We automatically annotate the error spans in the incorrect translations for 34514 samples out of 36476, by deterministically comparing the incorrect translation to either the good translation or the reference sentence. The automatic annotation methods mainly depends on the way the challenge sets for each phenomena were constructed, and contain only word-level annotations following the annotation guidelines. The details about the automatic annotation methods are as following:

Annotation of addition, omission and substitutions. This method tokenises the good translation and incorrect translation, and compares the tokens to annotate word-level addition, omission and substitutions which may occur multiple times. It is only used to annotate the simpler cases of substitutions, when each word was replaced with another word.

Annotation of substitution of a variable-sized span comparing to the correct translation. This method tokenises the good translation and the incorrect translation and then finds a single word-level error span with variable size.

Annotation of substitution of a variable-sized span comparing to the reference sentence. Similar to "Annotation of substitution of a variable-sized span comparing to the correct translation", this method tokenises the reference and the incorrect translation and then finds a single word-level error span with variable size.

Annotation of the date-time translation errors. In the Hallucination - Date-Time challenge set, the incorrect translations were built by substituting a month name in the reference with another month. This method finds the month names which are different in the incorrect translations and the reference, ignoring the months replaced with their corresponding abbreviations.

Annotation of the unit-conversion translation errors. In the Hallucination - Unit Conversion phenomenon, the unit mentions in the reference (e.g. 100 feet) were replaced with either the

wrong amount (e.g. 100 metres) or wrong unit (30.5 feet) in the incorrect translation. Using the Python package quantulum3[6], we detect the amount and units used in the incorrect translation, and annotate either the wrong amount or the wrong unit, according to the phenomenon category label (hallucination-unit-conversion-unit-matches-ref and hallucination-unit-conversion-amount-matches-ref respectively).

Annotation of the error where two words in the good translation were swapped. In ordering-mismatch challenge set, the incorrect sentence was generated by swapping the places of two words in the good translation. This method computes the annotations when two spans were swapped, and we manually annotated 4 samples which the method was not able to correctly annotate.

Annotation of the whole sentence. This method trivially annotates the whole incorrect translation as an error. For examples belonging to the following Mistranslation - Sentence-Level Meaning Error phenomena, constructed using the XNLI dataset, we automatically mark the entire sentence as an error: xnli-addition-contradiction, xnli-addition-neutral, xnli-omission-contradiction, xnli-omission-neutral. Despite some degree of lexical overlap between the good- and incorrect-translation, the incorrect-translation is drawn from either a contradiction or neutral hypothesis in the XNLI dataset, and will therefore by definition *not be a translation* of the premise (i.e. the sentence extracted as the good-translation).

4.2 Manual Annotation

Automated annotation is suitable for many of the examples, e.g. where the good and incorrect translations only exhibit differences relevant to the particular phenomenon indicated by the category label. However, it is not suitable in all cases, for example where the good and incorrect translations contain additional differences (not related to the error phenomenon), which could result in the automatic annotation method introducing annotation errors. We identified four phenomena for which automated annotation was unsuitable, and submitted all examples from these categories for manual annotation. The table below lists the four ACES phenomenon labels and their corresponding category in the manual annotation guidelines.

ACES Phenomenon Label	Category in Annotation Guidelines
coreference-based-on-commonsense	coreference
hallucination-real-data-vs-ref-word	hallucination
hallucination-real-data-vs-synonym	hallucination
lexical-overlap	word swap

We extracted a total of 2,006 examples belonging to these phenomena (427 hallucination, 559 coreference, and 1020 word swap), with examples for the following languages: English (471), French (551), German (456), Japanese (322), Korean (4), Marathi (44), and Russian (158). The manual annotation of these examples was completed by a team of seven annotators (one per language), who are either professional translators or linguists. The annotators were provided with a set of general guidelines plus specific instructions for each of the different phenomena listed above. The annotation guidelines are summarised in the following sections and the complete set of guidelines given to the annotators is provided in Appendix F: ACES Span Annotation Guidelines.

Automated checks were carried out over the manual annotations to provide a basic validation. These checks were used to ensure that 1) each example had been annotated, i.e. contained at least one span of text within tags, 2) all spans were marked with an open and close tag (i.e. the

---

6 https://github.com/nielstron/quantulum3

number of open and close tags per example, should match), 3) no changes had been made to the example text other than the addition of the tags. Examples that failed these checks were sent to the annotators for re-annotation. We also automatically identified and resolved instances where additional whitespace was introduced (in error) at the start or end of an error span, ensuring that the annotated text and original (unannotated) text differed only in terms of the presence/absence of error tags.

4.2.1 Overview of Annotation Guidelines. We split the annotation guidelines into a) general guidelines suitable for annotating all examples, and b) error type-specific guidelines intended for annotating specific categories. The annotators are presented with an ACES phenomenon label representing the type of error present, and two sentences: A and B, where B is the incorrect translation (i.e. contains one or more errors) and A is either the good translation or the reference (depending on the phenomenon). The annotators are asked to identify and mark *all* error spans in sentence B that belong to the error type indicated by the phenomenon label. Error spans are marked with tags (<>) at the word level, i.e. in the case that the error is a *misspelling* (e.g. "combuter" instead of "computer") the complete word (i.e. "combuter") should be marked.

    General guidelines. The general guidelines may be applied for the annotation of any example in ACES. We begin by defining four possible operations to mark error spans: *addition*, *substitution*, *deletion*, and *reordering* (see Table 2). In simple scenarios, a single operation may be sufficient to annotate an example. In more complex scenarios multiple operations may be required.

Addition: a text span that is not present in sentence A is included in sentence B

    Sentence A: The cat is a species of small carnivorous mammal.
    Sentence B: The cat is a **<domestic>** species of small carnivorous mammal.

Substitution: a text span in sentence A is substituted with a different text span in sentence B

    Sentence A: Female domestic cats can have kittens from spring to late autumn.
    Sentence B: Female domestic cats can have kittens from **<May>** to **<December>**.

Deletion: a text span that is present in sentence A is omitted from sentence B

    Sentence A: Feral cats are domestic cats that were born in or have reverted to a wild state.
    Sentence B: Feral cats are domestic cats **<>**or have reverted to a wild state.

Reordering: a text span in sentence A that appears in a different position in sentence B

    Sentence A: Montreal is the second most populous city in Canada and the most
              populous city in the province of Quebec.
    Sentence B: Montreal is the **<>**most populous city in Canada and the **<second>** most
              populous city in the province of Quebec.

Table 2
Manual annotation guidelines: Operations for general guidelines

    Error type-specific guidelines: Additionally, we include specific guidelines for the annotation of three phenomenon categories: *hallucination*, *coreference*, and *word swap* (see Table 3). The annotation of examples belonging to these categories may be achieved by marking the presence of one or more operations. For example, the hallucination example in Table 3 contains both an "addition" (i.e. **<Welsh, French,>**) and a "substitution" (i.e. Gaelic → **<Garlic>**). The three categories, for which we provide *error type-specific guidelines*, cover all of the examples submitted for manual annotation.

16

Hallucination: text that is not present in sentence A is observed in sentence B or a word in sentence A is replaced by a more frequent or *orthographically similar* word in sentence B

Sentence A: The official languages of Scotland are: English, Scots, and Scottish Gaelic.

Sentence B: The official languages of Scotland are: English, **<Welsh, French,>** Scots, and Scottish **<Garlic>**.

Coreference: a pronoun in sentence A is replaced with a (potentially) inappropriate noun-phrase in sentence B

Sentence A: The cat had caught the mouse and it was trying to wriggle free.

Sentence B: The cat had caught the mouse and **<the cat>** was trying to wriggle free.

Word swap: the position of a word or text span in sentence A appears swapped in sentence B

Sentence A: Their music is considered by many as an alternative metal with rap metal and industrial metal influences, which according to previous interviews call themselves "murder - rock".

Sentence B: Their music is considered by many as **<industrial>** metal with rap metal and **<alternative>** metal influences. According to previous interviews, they consider themselves "murder rock".

Table 3
Manual annotation guidelines: Error type-specific guidelines

4.2.2 Development of Manual Annotation Guidelines. To aid in the development and refinement of the annotation guidelines, we conducted a two-phase annotation pilot. In the first phase, we drew up the set of formal guidelines (described in Section 4.2.1). In the second phase, we verified the guidelines and measured inter-annotator agreement. We then asked professional annotators to complete the manual annotation of the four ACES phenomena listed above, using the guidelines.

In the first pilot phase, four of the authors of the paper[7] manually annotated error spans for a sample of 100 examples with English as the target language, randomly selected across all phenomena in ACES. The annotators had access to the source-language sentence, the three target-language translations: good- incorrect- and reference-translation, and the phenomenon label. We considered only the target language side and marked one or more error spans in the incorrect translation only. We then conducted an adjudication exercise in which all four annotators manually compared the four sets of annotations for each example and discussed our strategies for annotation. From this, we derived a set of general guidelines to accommodate the annotation of any example in ACES. We then added specific guidelines for examples belonging to the categories: *hallucination*, *coreference*, and *word swap*.

In the second pilot phase, we verified the quality of the manual annotation guidelines. To verify the general guidelines, and provide a gold standard against which to measure the automated evaluation method, the same four annotators from the first pilot phase annotated another sample of 100 examples with English as the target language, randomly selected across all ACES phenomena. To verify the quality of the span annotations, we automatically measured inter-annotator agreement. We computed the percentage of exact matches[8] as total_exact_matches

---

7 Two annotators for the first pilot phase are native English speakers; two are fluent English speakers
8 We ignore both leading and trailing whitespace when comparing spans

divided by total_spans_marked, i.e. where all four annotators agree on the same error span, as 81.82% (examples=100, total spans=110, exact-match spans=90), indicating high agreement[9]. We also verified the type-specific guidelines for annotating *hallucination*, *coreference*, and *word swap*. As the *coreference* category requires manual annotation in German (ACES contains only en-de examples for the *coreference-based-on-commonsense* phenomenon), and examples of the other phenomena exist for English, we asked two native German / fluent English speakers[10] to annotate a randomly selected sample of 100 examples (25 examples from each of the relevant ACES phenomenon categories). We report inter-annotator agreement of 77.40% (examples=100, total spans=146, exact-match spans=113).

In addition to measuring inter-annotator agreement, we also examined the examples where two or more annotators marked different spans. We concluded that the majority of differences arose from simple human errors as opposed to differing interpretations of the guidelines. For example, annotators sometimes accidentally marked longer spans than necessary, or marked the presence of a deletion in the wrong position. We concluded that many of these mistakes could have been avoided had the annotators carefully double-checked their annotations. We therefore added a note to the guidelines to this effect, but made no further changes to the instructions. It is also worth noting that for a handful of examples, the presence of Machine Translation led to annotators struggling to agree on a correct annotation – an issue that is not easily resolved, but is infrequent in the ACES dataset.

5. Evaluation Methodology

Table 4 lists the baseline, reference-based, and reference-free metrics from WMT 2022 and 2023 that provide segment-level judgements and cover all of the language pairs in ACES. We indicate whether metrics are embeddings-based with a subset of metrics using the supervision signal provided by Direct Assessment (DA) judgements from WMT (Bojar et al. 2016) or MQM (Lommel, Burchardt, and Uszkoreit 2014) annotations, LLM-based, or rely on surface-level overlap with the reference.

We briefly summarise the metrics here, grouping them into broad categories based on their design characteristics. The metrics that ***rely on surface overlap with the reference*** include several baseline metrics: **BLEU** (Papineni et al. 2002), **chrF** (Popović 2017) and the **spBLEU** (Goyal et al. 2022) metrics F101SPBLEU and F200SPBLEU, for which the SentencePiece tokeniser (Kudo and Richardson 2018) was trained using data from the FLORES-101 or -200 languages respectively. It also includes the 2023 participant metrics based on F-scores and inspired by chrF++: **Tokengram_F** and **Partokengram_F** (Dréano, Molloy, and Murphy 2023b).

The largest group is ***embedding-based metrics***. Many are based on the **COMET** architecture: **COMET-20** and **COMET-QE** (Rei et al. 2020), Unbabel's WMT 2022 submission **COMET-22** (Rei et al. 2022), and Microsoft's WMT 2022 submissions **MS-COMET-22** and **MS-COMET-QE-22** (Kocmi, Matsushita, and Federmann 2022). The **XCOMET** family of metrics, trained to identify errors in sentences along with a final quality score, includes **XCOMET-XL**, **XCOMET-XXL**, and **XCOMET-QE**, and the two ensemble metrics: **XCOMET-Ensemble** and **XCOMET-QE-Ensemble**. The **COMET-Kiwi** (Rei et al. 2022) metric and the **COMETKiwi-XL** and **COMETKiwi-XXL** metrics from 2023 form another family. The **Cometoid22** (Gowda, Kocmi, and Junczys-Dowmunt 2023) student metrics are trained to mimic teacher scores from COMET-22 without access to the reference. (The suffix [WMT-21,22,23] indicates the training data cut-off

---

9 Highest inter-annotator agreement with three annotators: 90.48% (examples=100, total spans=105, exact-match spans=95)

10 One annotator for the second pilot phase was also an author of this paper

year.) The remaining metrics are based on a range of different architectures: **BERTScore** (Zhang et al. 2020), **BLEURT20** (Sellam et al. 2020), **YiSi-1** (Lo 2019), **UniTE** (Wan et al. 2022a), **MATESE** and **MATESE-QE** (Perrella et al. 2022b), **REUSE** (Mukherjee and Shrivastava 2022), **eBLEU** (ElNokrashy and Kocmi 2023), and **XLsim** (Mukherjee and Shrivastava 2023). The **MetricX** family includes the **metricx_*_DA** and **metricx_*_MQM** metrics from 2022 and **MetricX-23** and **MetricX-23-QE** (Juraska et al. 2023) from 2023. The Huawei metrics include **Cross-QE**, **HWTSC-Teacher-Sim**, and **HWTSC-TLM** (Liu et al. 2022), and **KG-BERTScore** (Liu et al. 2022; Wu et al. 2023) which incorporates a multilingual knowledge graph.

The ***LLM-based*** metrics group comprises two WMT 2023 metrics: **Embed_Llama** (Dréano, Molloy, and Murphy 2023a) which uses pre-trained LLaMA2 embeddings without finetuning, and **GEMBA-MQM** (Kocmi and Federmann 2023a) – a GPT-based metric for error quality span marking. Finally, **Random-sysname** is a random baseline which samples scores from a Gaussian distribution based on random mean value. It was included in 2023 to provide a context to scores and also to detect errors in metric meta-evaluations. In addition to these metrics, we also conducted some experiments on using LLMs for evaluation as listed below.

5.1 LLM Metrics

Following the rapid adoption of LLM-based approaches to address a range of NLP tasks, there has also been a steady increase in the use of LLMs for MT evaluation with apparently promising results (Xu et al. 2023; Lu et al. 2023; Kocmi and Federmann 2023a). We note that these observations are often limited to system-level evaluation and also to high-resource language pairs. We thus intend to investigate the extent to which these LLMs can be used for MT evaluation more holistically through the ACES dataset.

We consider three variants of using LLMs for evaluation. The first one is GEMBA-DA (Kocmi and Federmann 2023a) where the model (GPT Davinci-003 a predecessor to GPT-4 model) is prompted using a zero-shot approach to produce a translation score between 0 and 100. Note that GEMBA-DA was the precursor of the GEMBA-MQM model, which was discussed previously. For the next two methods, we considered LLaMA2 (7B) (Touvron et al. 2023) and FLAN-ALPACA-XL (Chia et al. 2023) (3B) which is Flan-T5 (Chung et al. 2022) fine-tuned on the Alpaca dataset (Taori et al. 2023). We chose LLaMA2 (7B), despite it being predominantly trained in English, to see if the accidental multilingual tokens are enough to provide multilingual evaluation. We included FLAN-ALPACA-XL as it is a *smaller* LLM and the base model was trained with multilingual data. [11]

For these LLMs (FLAN-ALPACA-XL and LLaMA2), we experimented with both zero-shot and five-shot prompting. In five-shot prompting, five examples of scored translations across varying scoring ranges and language pairs were provided with the prompt. However, we found that five-shot prompting performed poorly in our initial experiments and therefore we report only the zero-shot results. We provide the prompt templates in the Appendix. For the postprocessing of outputs from the above LLMs, we included the first rational number that appeared in the output from the respective models as the *score* produced by that LLM. In the scenario in which no number was found, the example was given a score of 0. In such examples, the overgenerated text generally consisted of a hallucinated example of a source-reference-translation triplet.

As ACES is a contrastive dataset, we also experimented with providing a prompt that compares the two translations, labelled A and B respectively, and instructs the LLM to select the *better* translation. However, in our initial experiments, we found that the models typically

---

11 We also conducted experiments on BLOOM (Scao et al. 2022) but found the majority of outputs produced by the BLOOM-7B model to be unintelligible which could not be converted into scores

produce an option followed by the generation of both of the candidate translations. This copying of translations makes it hard to identify if the generation of the option was a result of the model actually performing the evaluation or an artefact of the overgeneration.

5.2 Metrics with error spans

In addition to the above metrics, we also conduct baseline experiments for SPAN-ACES. We include recently developed metrics that directly predict error spans while generating the scores, namely XCOMET-XL (Guerreiro et al. 2023) and GEMBA-MQM (Kocmi and Federmann 2023a). These metrics also provide severity of the error for the predicted error span - minor, major, and critical.

Additionally, we derive baselines from existing metrics that were trained to only produce scores. We re-purpose the work in Rei et al. (2023), which included the proposal of several neural explainability methods for interpreting state-of-the-art fine-tuned neural machine translation metrics such as COMET and UNITE. In one of these techniques, *embed–align*, they calculate the maximum cosine similarity between each translation token embedding and the reference and/or source token embeddings (Tao et al. 2022) and assign that scalar value to each translation token. Starting from embed-align scores attributed to each translation token, we generate error spans over the translations by marking any token which has an embed-align score higher than a constant threshold. We set the threshold that yields the span predictions with the highest Recall@K score on the WMT 2021 MQM annotations development dataset[12]. This method produces six different types of span predictions: embed–align[mt, src], embed–align[mt, ref] and embed–align[mt, src; ref] using the embeddings extracted from each of the COMET-22 and UNITE models [13].

5.3 Evaluation of Metrics

For all phenomena in ACES where we generated more than 1,000 examples, we randomly subsample 1,000 examples according to the per language pair distribution to include in the final challenge set to keep the evaluation of new metrics tractable.

We follow the evaluation of the challenge sets from the 2021 edition of the WMT metrics shared task (Freitag et al. 2021b) and report performance with Kendall's tau-like correlation[14]. The Kendall's tau-like metric (see Equation 1) measures the number of times a metric scores the good translation above the incorrect translation (concordant) and equal to or lower than the incorrect translation (discordant). Ties are considered as discordant. Note that a higher τ indicates a better performance and that the values can range between -1 and 1.

$$\tau = \frac{concordant - discordant}{concordant + discordant} \tag{1}$$

We discuss the evaluation on SPAN-ACES closer to its results section.

	supervised	surface overlap	base-embedding	LLM-based	2022	2023
BLEU		✓			✓	✓
f101spBLEU		✓			✓	
f200spBLEU		✓			✓	✓
chrF		✓			✓	✓
BERTScore			?		✓	✓
BLEURT20	WMT human eval		BERT		✓	✓
COMET-20			XML-R		✓	
COMET-QE			XML-R?		✓	
YiSi-1			?		✓	✓
Random-sysname						✓
COMET-22*†	DA+MQM				✓	✓
MATESE	MQM				✓	
metricx_xl_DA_2019	DA		mt5		✓	
metricx_xl_MQM_2020	MQM		mt5		✓	
metricx_xxl_DA_2019	DA		mt5		✓	
metricx_xxl_MQM_2020	MQM		mt5		✓	
MS-COMET-22	human judgements		mt5		✓	
UniTE					✓	
UniTE-ref †					✓	
eBLEU						✓
embed_llama			Llama 2	✓		✓
MetricX-23	DA+MQM		mT5			✓
MetricX-23-b	DA+MQM		mT5			✓
MetricX-23-c	DA+MQM		mT5			✓
partokengram_F		✓?				✓
tokengram_F		✓				✓
XCOMET-Ensemble	DA+MQM		XLM-R			✓
XCOMET-XL †	DA+MQM		XLM-R			✓
XCOMET-XXL	DA+MQM		XLM-R			✓
XLsim	WMT human eval		XLM-R			✓
COMETKiwi*	DA		InfoXLM		✓	✓
Cross-QE			?		✓	
HWTSC-Teacher-Sim			paraphrase-multilingual -mpnet-base-v2		✓	
HWTSC-TLM			?		✓	
KG-BERTScore					✓	✓
MATESE-QE	MQM				✓	
MS-COMET-QE-22*					✓	✓
REUSE			BERT		✓	
UniTE-src					✓	
cometoid22-wmt21	?		InfoXLM			✓
cometoid22-wmt22	?		InfoXLM			✓
cometoid22-wmt23	?		InfoXLM			✓
CometKiwi-XL			XLM-R			✓
CometKiwi-XXL			XLM-R			✓
GEMBA-MQM †				✓		✓
MetricX-23-QE	DA+MQM		mT5			✓
MetricX-23-QE-b	DA+MQM		mT5			✓
MetricX-23-QE-c	DA+MQM		mT5			✓
XCOMET-QE-Ensemble	DA+MQM		XLM-R			✓
XLsimQE	WMT human eval		XLM-R			✓

Table 4
Basline (top), reference-based (middle), and reference-free (bottom) metrics from WMT 2022 and 2023
Metrics shared tasks. * denotes a participating metric from 2022 that was used as a baseline in 2023. †
denotes that metrics were used as baselines for SPAN-ACES. ? indicates no information was made available.

Examples	addition	omission	mistrans.	untranslated	do not translate	overtrans.	undertrans.	real-world knowledge	wrong language	punctuation	ACES-Score
	999	999	24457	1300	100	1000	1000	2948	2000	1673	
BLEU	0.748	0.435	-0.229	0.353	0.600	-0.838	-0.856	-0.768	0.661	0.638	-2.7
f101spBLEU	0.662	0.590	-0.084	0.660	0.940	-0.738	-0.826	-0.405	0.638	0.639	-0.1
f200spBLEU	0.664	0.590	-0.082	0.687	0.920	-0.752	-0.794	-0.394	0.658	0.648	0.1
chrF	0.642	0.784	0.162	0.781	0.960	-0.696	-0.592	-0.294	0.691	0.743	3.7
BERTScore	0.880	0.750	0.320	0.767	0.960	-0.110	-0.190	0.031	0.563	0.849	10.6
BLEURT-20	0.437	0.810	0.429	0.748	0.860	0.200	0.014	0.401	0.533	0.649	12.0
COMET-20	0.437	0.808	0.378	0.748	0.900	0.314	0.112	0.267	0.033	0.706	12.2
COMET-QE	-0.538	0.397	0.378	0.135	0.120	0.622	0.442	0.322	-0.505	0.251	6.6
YiSi-1	0.770	0.866	0.356	0.730	0.920	-0.062	-0.076	0.110	0.431	0.734	11.5
COMET-22	0.333	0.806	0.566	0.536	0.900	0.690	0.538	0.574	-0.318	0.539	16.4
metricx_xl_DA_2019	0.395	0.852	0.545	0.722	0.940	0.692	0.376	0.740	0.521	0.670	17.2
metricx_xl_MQM_2020	-0.281	0.670	0.523	0.579	-0.740	0.718	0.602	0.705	-0.126	0.445	13.1
metricx_xxl_DA_2019	0.303	0.832	0.580	0.762	0.920	0.572	0.246	0.691	0.250	0.630	15.3
metricx_xxl_MQM_2020	-0.099	0.534	0.578	0.651	0.880	0.752	0.552	0.712	-0.321	0.369	13.5
MS-COMET-22	-0.219	0.686	0.397	0.504	0.700	0.548	0.290	0.230	0.041	0.508	10.0
UniTE	0.439	0.876	0.501	0.571	0.920	0.496	0.302	0.624	-0.337	0.793	14.9
UniTE-ref	0.359	0.868	0.535	0.412	0.840	0.640	0.398	0.585	-0.387	0.709	15.5
COMETKiwi	0.361	0.830	0.631	0.230	0.780	0.738	0.574	0.582	-0.359	0.490	16.9
Cross-QE	0.163	0.876	0.546	-0.094	0.320	0.726	0.506	0.446	-0.374	0.455	14.4
HWTSC-Teacher-Sim	-0.031	0.495	0.406	-0.269	0.700	0.552	0.456	0.261	-0.021	0.271	10.1
HWTSC-TLM	-0.363	0.345	0.384	0.154	-0.040	0.544	0.474	0.071	-0.168	0.634	7.0
KG-BERTScore	0.790	0.812	0.489	-0.456	0.760	0.654	0.528	0.487	0.306	0.255	17.5
MS-COMET-QE-22	-0.177	0.678	0.439	0.388	0.240	0.518	0.386	0.248	-0.197	0.523	9.9
UniTE-src	0.285	0.930	0.599	-0.615	0.860	0.698	0.540	0.537	-0.417	0.733	15.7
Average	0.290	0.713	0.389	0.404	0.735	0.312	0.167	0.282	0.075	0.578	10.9

Table 5

2022 Results. Average Kendall's tau-like correlation results for the nine top level categories in the ACES ontology, plus the additional fluency category: punctuation. The horizontal lines delimit baseline metrics (top), participating reference-based metrics (middle) and participating reference-free metrics (bottom). The best result for each category is denoted by bold text with a green highlight. Note that *Average* is an average over averages. The last column shows the ACES-Score, a weighted sum of the correlations. The ACES-Score ranges from –29.1 (all phenomena have a correlation of -1) to 29.1 (all phenomena have a correlation of +1).

	addition	omission	mistrans.	untranslated	do not translate	overtrans.	undertrans.	real-world knowledge	wrong language	punctuation	ACES-Score
Examples	*999*	*999*	*24457*	*1300*	*100*	*1000*	*1000*	*2948*	*2000*	*1673*	
BERTscore	**0.872**	0.754	0.318	0.771	0.940	-0.186	-0.288	0.030	0.551	**0.844**	9.7
BLEU	0.742	0.427	-0.227	0.353	0.580	-0.838	-0.856	-0.768	0.660	0.704	-2.8
BLEURT-20	0.435	0.812	0.427	0.743	0.860	0.202	0.014	0.388	0.536	0.708	12.0
chrF	0.644	0.784	0.162	**0.781**	**0.960**	-0.696	-0.592	-0.294	0.693	0.773	3.7
COMET-22	0.295	0.822	0.402	0.718	0.820	0.502	0.258	0.382	0.078	0.673	13.458
CometKiwi	0.536	**0.918**	0.614	-0.105	0.520	0.766	0.604	0.577	-0.307	0.765	**17.9**
f200spBLEU	0.666	0.584	-0.082	0.680	0.920	-0.752	-0.794	-0.394	0.657	0.708	0.041
MS-COMET-QE-22	-0.179	0.674	0.440	0.394	0.300	0.524	0.382	0.262	-0.195	0.632	10.0
Random-sysname	-0.117	-0.117	-0.116	-0.083	-0.100	-0.118	-0.152	-0.245	-0.113	-0.074	-3.6
YiSi-1	0.766	0.868	0.354	0.720	0.940	-0.062	-0.076	0.110	0.421	0.763	11.5
eBLEU	0.674	0.682	0.197	0.739	0.880	-0.662	-0.684	-0.042	**0.771**	0.270	3.4
embed_llama	0.211	0.457	0.016	0.503	0.400	-0.170	-0.492	-0.165	0.154	0.476	1.054
MetricX-23	-0.027	0.568	0.578	0.473	0.800	0.790	0.586	0.766	-0.486	0.636	14.1
MetricX-23-b	-0.135	0.622	0.572	0.613	0.860	0.772	0.568	0.749	-0.444	0.532	13.8
MetricX-23-c	-0.015	0.794	0.617	0.611	0.800	0.740	0.526	**0.783**	-0.629	0.527	15.0
partokengram_F	0.087	0.191	-0.034	0.310	0.140	-0.042	-0.028	0.032	0.508	0.171	1.9
tokengram_F	0.698	0.758	0.160	0.779	**0.960**	-0.732	-0.632	-0.273	0.687	0.830	3.5
XCOMET-Ensemble	0.311	0.786	0.663	0.379	0.780	**0.794**	0.612	0.708	-0.423	0.595	17.3
XCOMET-XL	0.169	0.542	0.570	0.222	0.800	0.656	0.464	0.582	-0.367	0.220	13.3
XCOMET-XXL	-0.119	0.413	0.547	0.234	0.600	0.736	0.568	0.508	-0.507	0.509	11.6
XLsim	0.429	0.618	0.153	0.643	0.820	-0.210	-0.290	-0.044	0.392	0.753	5.4
cometoid22-wmt21	-0.339	0.658	0.493	-0.076	0.280	0.670	0.566	0.362	-0.454	0.608	10.4
cometoid22-wmt22	-0.301	0.674	0.493	-0.119	0.280	0.686	0.538	0.340	-0.472	0.599	10.534
cometoid22-wmt23	-0.253	0.702	0.502	-0.046	0.420	0.750	0.590	0.362	-0.319	0.557	11.9
CometKiwi-XL	0.239	0.828	0.624	0.239	0.440	0.762	0.560	0.563	-0.380	0.630	16.0
CometKiwi-XXL	0.361	0.828	0.653	0.414	0.320	0.774	0.560	0.683	-0.537	0.503	16.8
GEMBA-MQM	0.037	0.281	0.153	0.094	0.140	0.466	0.276	0.268	-0.150	0.015	6.4
KG-BERTScore	0.538	0.912	0.585	-0.206	0.700	0.772	0.606	0.594	-0.307	0.654	**18.0**
MetricX-23-QE	0.045	0.678	0.654	0.379	0.460	0.772	0.612	0.654	-0.702	0.226	14.6
MetricX-23-QE-b	0.027	0.760	0.663	0.489	0.480	0.758	**0.620**	0.647	-0.673	0.256	15.1
MetricX-23-QE-c	-0.115	0.664	**0.721**	0.384	0.340	0.726	0.618	0.753	-0.712	0.375	13.8
XCOMET-QE-Ensemble	0.277	0.754	0.644	0.181	0.720	0.764	0.582	0.626	-0.519	0.449	16.1
XLsimQE	0.205	0.383	0.087	-0.694	0.940	0.454	0.352	0.042	0.307	0.671	8.0
Average	0.232	0.639	0.382	0.349	0.609	0.314	0.187	0.289	-0.069	0.532	10.0

Table 6
2023 Results. Average Kendall's tau-like correlation results for the ACES top-level categories and ACES-Scores (final column). Metrics are grouped into baseline (top), and participating reference-based (middle) and reference-free (bottom) metrics. Note that *Average* is an average over averages. Best results are highlighted in green.

6. Results

6.1 Phenomena-level Results

We begin by providing a broad overview of metric performance on the different phenomena categories, before conducting more detailed analyses in Section 7. We restrict the overview to the metrics which provide a) segment-level scores and b) scores for all language pairs and directions in ACES. After filtering according to these criteria, 24 metrics from 2022 remain: nine baseline, eight reference-based, and seven reference-free metrics. In 2023, 33 metrics fulfil these criteria: 10 baseline, 11 reference-based, and 12 reference-free metrics.

We first calculate Kendall's tau-like correlation scores for all of the ACES examples (see Equation 1). We then report the average score over all examples belonging to each of the nine top-level accuracy categories in ACES, plus the fluency category *punctuation* (see Tables 5 and 6). In addition, we calculate the ACES-Score, a weighted combination of the top-level categories, which allows us to identify high-level performance trends of the metrics (see Equation 2). The weights correspond to the values under the MQM framework (Freitag et al. 2021a) for major (weight=5), minor (weight=1) and fluency/punctuation errors (weight=0.1). We categorise untranslated, do not translate and wrong language as minor errors due to the ease with which they can be identified with automatic language detection tools or during post-editing. We also include real-world knowledge under minor errors since we do not generally expect MT evaluation metrics to have any notion of real-world knowledge and do not wish to punish them for this. Note that the ACES-Score ranges from -29.1 (all phenomena have a correlation of -1) to 29.1 (all phenomena have a correlation of +1).

$$\text{ACES} = sum \left\{ \begin{array}{c} 5 * \tau_{\text{addition}} \\ 5 * \tau_{\text{omission}} \\ 5 * \tau_{\text{mistranslation}} \\ 1 * \tau_{\text{untranslated}} \\ 1 * \tau_{\text{do not translate}} \\ 5 * \tau_{\text{overtranslation}} \\ 5 * \tau_{\text{undertranslation}} \\ 1 * \tau_{\text{real-world knowledge}} \\ 1 * \tau_{\text{wrong language}} \\ 0.1 * \tau_{\text{punctuation}} \end{array} \right\} \quad (2)$$

We report an overview of the results for WMT 2022 in Table 5 and the results for WMT 2023 in Table 6. Using the ACES-Score (the final column in each of the tables) we can see at a glance that the majority of the metrics submitted to the WMT 2022 shared task outperform the baseline metrics. The same is true of the WMT 2023 metrics – except for CometKiwi, a successful submission from 2022 which was used as a baseline in 2023 – the majority of the 2023 baseline

---

12 threshold=0.1 for COMET-22, threshold=0.14 for UNiTE
13 We use the wmt22-comet-da version for COMET-22 and SRC+REF version for UNiTE
14 Evaluation scripts are available here: https://github.com/EdinburghNLP/ACES

metrics are outperformed by the metrics submitted by participants. Interestingly, in both years, many reference-free metrics performed on par with reference-based metrics. This is because our challenge sets are constructed to make the reference useless (ambiguous translation, discourse connectives, *etc.,*), or misleading (hallucinations, lexical overlap, sentence-level meaning error). Note that we cannot directly compare the results from 2022 and 2023 – for a small subset (2,659; approx. 7%) of the ACES examples different results were returned in 2022 and 2023 for metrics where no changes had been made (e.g. baseline metrics such as BLEU or CometKiwi, etc.)[15].

The best-performing metric in 2022 is a reference-free metric, namely KG-BERTSCORE, closely followed by the reference-based metric METRICX_XL_DA_2019. The best-performing metrics in 2023 are COMETKIWI (a reference-free baseline metric), and KG-BERTSCORE. Perhaps unsurprisingly, BLEU is one of the worst performing metrics, underperformed only by the random baseline, RANDOM-SYSNAME, in 2023. We caution that we developed ACES to investigate strengths and weaknesses of metrics on a phenomena level – hence, we advise the reader not to draw any conclusions based solely on the ACES-Score.

Our observations regarding the metric performance were similar for both 2022 and 2023, and the following three points hold true for both years. Firstly, we observed that metric performance varies greatly and there is no clear winner in terms of performance across all of the categories. There is also a high degree of variation in terms of metric performance when each category is considered in isolation. Secondly, whilst each of the categories proves challenging for at least one metric, some categories are more challenging than others. For example, looking at the average scores in the last row of Table 5, and without taking outliers into account, we might conclude that addition, undertranslation, real-world knowledge, and wrong language (all with average Kendall tau-like correlation of < 0.3) present more of a challenge than the other categories. On the other hand, for omission and do not translate (with an average Kendall tau-like correlation of > 0.7 in 2022 and > 0.6 in 2023) metric performance is generally rather high.

Thirdly, we also observe variation in terms of the performance of metrics belonging to the baseline, reference-based, and reference-free groups. For example, in both years, the baseline metrics generally appear to struggle more on the overtranslation and undertranslation categories than the metrics belonging to the other groups. Reference-based metrics also appear to perform better overall on the untranslated category than the reference-free metrics. This makes sense as a comparison with the reference is likely to highlight tokens that ought to have been translated.

Whilst there are many similarities in the performance trends observed in 2022 and 2023, there are also some differences. In 2023, we observe that the reference-free group exhibits overall stronger performance compared with the other groups, but in particular for the *mistranslation*, *overtranslation*, *undertranslation*, and *real-world knowledge* categories. We observe that (unlike in 2022) some of the 2023 metrics perform similarly to or worse than the baseline metrics. In particular, EMBED_LLAMA and GEMBA-MQM which are designed using Large Language Models (LLMs), struggle with this challenge set. This suggests that we need better design strategies in using the rich representations from LLMs for MT evaluation.

6.2 Mistranslation Results

Next, we drill down to the fine-grained categories of the largest category: *mistranslation*. We present metric performance on its sub-level categories (*discourse*, *hallucination*, and *other*) in Table 7 (2022 results) and Table 8 (2023 results). The *discourse* sub-category includes errors

---

-	disco.	halluci.	other
Examples	*3698*	*10270*	*10489*
BLEU	-0.048	-0.420	-0.251
f101spBLEU	0.105	-0.206	-0.153
f200spBLEU	0.094	-0.191	-0.149
chrF	0.405	-0.137	0.161
BERTScore	0.567	-0.058	0.362
BLEURT-20	0.695	0.142	0.402
COMET-20	0.641	0.016	0.399
COMET-QE	0.666	0.303	0.208
YiSi-1	0.609	0.019	0.368
COMET-22	0.682	0.461	0.542
metricx_xl_DA_2019	0.701	0.493	0.458
metricx_xl_MQM_2020	0.573	0.677	0.394
metricx_xxl_DA_2019	0.768	0.541	0.463
metricx_xxl_MQM_2020	0.716	**0.713**	0.392
MS-COMET-22	0.645	0.148	0.360
UniTE	0.746	0.322	0.424
UniTE-ref	**0.776**	0.396	0.437
COMETKiwi	0.733	0.493	**0.637**
Cross-QE	0.644	0.395	0.563
HWTSC-Teacher-Sim	0.594	0.296	0.330
HWTSC-TLM	0.756	0.306	0.151
KG-BERTScore	0.593	0.387	0.472
MS-COMET-QE-22	0.626	0.243	0.416
UniTE-src	0.772	0.463	0.551
Average	0.586	0.242	0.331

Table 7
2022 Results. Average Kendall's tau-like correlation results for the sub-level categories in mistranslation: **disco**urse-level, **halluci**nation, and **other** errors. The horizontal lines delimit baseline metrics (top), participating reference-based metrics (middle) and participating reference-free metrics (bottom). The best result for each category is denoted by bold text with a green highlight. Note that *Average* is an average over averages.

	disco.	halluci.	other
Examples	*3698*	*10270*	*10489*
BERTscore	0.563	-0.062	0.361
BLEU	-0.042	-0.418	-0.250
BLEURT-20	0.695	0.141	0.398
chrF	0.406	-0.138	0.160
COMET-22	0.657	0.113	0.383
CometKiwi	0.779	0.465	0.580
f200spBLEU	0.095	-0.190	-0.150
MS-COMET-QE-22	0.631	0.240	0.417
Random-sysname	-0.117	-0.122	-0.111
YiSi-1	0.608	0.017	0.366
eBLEU	0.374	-0.166	0.282
embed_llama	-0.089	-0.140	0.189
MetricX-23	0.757	0.663	0.393
MetricX-23-b	0.749	0.656	0.390
MetricX-23-c	0.694	**0.755**	0.477
partokengram_F	-0.062	-0.101	0.027
tokengram_F	0.396	-0.132	0.157
XCOMET-Ensemble	**0.791**	0.566	0.626
XCOMET-XL	0.706	0.482	0.521
XCOMET-XXL	0.609	0.540	0.504
XLsim	0.217	-0.066	0.236
cometoid22-wmt21	0.782	0.286	0.400
cometoid22-wmt22	0.748	0.290	0.423
cometoid22-wmt23	0.758	0.223	0.478
CometKiwi-XL	0.752	0.501	0.602
CometKiwi-XXL	0.735	0.535	0.661
GEMBA-MQM	0.076	0.291	0.127
KG-BERTScore	0.685	0.466	0.580
MetricX-23-QE	0.728	0.604	0.628
MetricX-23-QE-b	0.694	0.617	0.666
MetricX-23-QE-c	0.747	0.659	**0.739**
XCOMET-QE-Ensemble	0.702	0.558	0.651
XLsimQE	0.053	0.050	0.134
Average	0.511	0.248	0.365

Table 8
2023 Results. Average Kendall's tau-like correlation results for the sub-level categories in mistranslation: **disco**urse-level, **halluci**nation, and **other** errors. The horizontal lines delimit baseline metrics (top), participating reference-based metrics (middle) and participating reference-free metrics (bottom). The best result for each category is denoted by bold text with a green highlight. Note that *Average* is an average over averages.

involving the mistranslation of discourse-level phenomena such as pronouns and discourse connectives. *Hallucination* includes errors at the word level that could occur due to hallucination by an MT model, for example, the use of wrong units, dates, times, numbers or named entities, as well as hallucinations at the subword level that result in nonsensical words. The *other* cub-category covers all other categories of mistranslation errors including overly literal translations of idioms and the introduction of ambiguities in the translation output.

As for the results overview in Section 6.1, we find that performance on the different sub-categories is variable, with no clear winner among the metrics in either 2022 or 2023. The results from both years suggest that hallucination phenomena are generally more challenging than discourse-level phenomena. Performance on the hallucination sub-category is poor overall,

	GEMBA-DA		LLAMA-2 (7B)		FLAN-T5-XL + Alpaca (3B)	
	REF	QE	REF	QE	REF	QE
addition	-0.235	-0.794	-0.607	-0.587	-0.834	-0.922
mistranslation	-0.031	-0.322	-0.58	-0.552	-0.656	-0.832
real-world knowledge	0.366	0.157	-0.58	-0.6	-0.280	-0.739
untranslated	-0.334	-0.606	-0.650	-0.626	-0.529	-0.631
do not translate	-0.100	-0.840	-0.64	-0.52	-0.180	-0.500
undertranslation	0.090	-0.286	-0.602	-0.602	0.016	-0.730
overtranslation	0.472	-0.034	-0.564	-0.524	0.026	-0.744
omission	-0.281	-0.568	-0.549	-0.503	-0.848	-0.854
punctuation	-0.306	-0.355	-0.646	-0.650	-0.875	-0.924
wrong language	0.026	-0.688	-0.55	-0.483	-0.632	-0.705
ACES-Score	-0.02	-12.0	-16.9	-16.1	-13.2	-23.1

Table 9
LLM results across three LLMs: GPT-4 through GEMBA-DA, LLAMA-2, and FLAN-T5-XL fine-tuned with Alpaca. REF: Reference based, QE: Quality Estimation/Reference-free. Using zero-shot prompting on LLMs for MT evaluation has results poorer than the surface overlap baselines in Table 5. This result worsens when the LLMs operate in a QE setting.

although it appears to be particularly challenging for the baseline metrics. We present additional, more fine-grained, performance analyses for individual phenomena in Section 7.

6.3 LLM Results

We report the results of the LLM experiments described in Section 5.1 in Table 9. Overall, we find MT evaluation via LLMs a hard task in the zero-shot setup. This is also evident in the results in Section 6.1 where we highlight the relatively low performance of GEMBA-MQM and EMBED-LLAMA. This is contrary to findings where LLMs show promising trends for evaluation at the system level (Fernandes et al. 2023; Kocmi and Federmann 2023b).

We find that of the three LLMs, GEMBA-DA has better (though still poor) performance. These results worsen for the reference-less setting where most of the phenomena have a negative correlation. Despite the instructions for DA scores to be assigned using a continuous scale of 0–100, we find that the LLMs tend to produce a peaked distribution. For example, GEMBA-DA produces only seven different scores for the full set of examples. This results in a higher number of ties which get penalised in Equation 1. Even after instructing the LLMs to output scores within the range of 0–100, we observed instances where the LLMs produced scores beyond that range.

These results suggest that while LLMs may perform well for MT evaluation under a specific setup like high-resource pairs or system-level evaluation, their zero-shot inference abilities for MT evaluation at segment-level are far from perfect. This can be attributed to a lack of multilingual training data (Kocmi and Federmann 2023a) as well as a limited numerical understanding of LLMs (Dziri et al. 2023). We additionally express concerns over *test-data leakage* as ACES is built on several other academic datasets (see Section 6.1) that may have been a part of the LLM training data (Carlini et al. 2020). We also note that these models are quite slow at inference. It takes approximately six hours to make a pass over the entire dataset using FLAN-T5-XL on a 24GB GPU, while it takes five days with two 24GB GPUs for LLAMA2 on 8bit precision.

6.4 Span-based Results

We first discuss the evaluation for SPAN-ACES and then report the results for the baseline methods discussed in Section 5.2.

6.4.1 Metrics for SPAN-ACES. We consider two different types of evaluation for SPAN-ACES extraction and contrastive evaluation:

Span Extraction: We first measure how well the methods that produce spans perform the task of identifying erroneous span(s) in a translation. We evaluate the predicted spans for the incorrect translation against the gold annotation. We calculate sample F1 and average across the dataset where a span is considered to be a true positive if the span exactly matches its ground truth, denoted as *Span-F1*. We also experimented with using partial matches between the gold error span and the predicted error span. However, using standardised tokenization based on words/sub-words/characters and then developing a threshold for partial match is not trivial and results in incorrect inflation of scores.

Contrastive Evaluation: To evaluate these methods on ACES and compare their results, we obtain span predictions for the good translation as well. We use a length heuristic where we measure the number of times the metric produces fewer spans for the good translation compared with the incorrect translation (concordant) and greater than or equal to the incorrect translation (discordant) and calculate the correlation as described in Section 5.3. If the severity of errors for the predicted spans is available, a use a weighted score based on the severity label.

6.4.2 Results. We now report the results of different models that produce error spans (and occasionally labels) from Section 6.4.1 on the SPAN-ACES dataset in Table 10. Overall, we find that these methods perform poorly on both the error span extraction and contrastive evaluation tasks.

On the span extraction task, we find that the derived methods – COMET-22 and UNITE – i.e. using attention maps over the source/reference sentences lead to higher Span-F1 scores than either XCOMET and GEMBA-MQM which were specifically designed to generate error spans. This adds some more evidence to the findings in Rei et al. (2023) that suggest metrics (COMET-22 and UNITE) tend to use token-level information that can be associated with tangible translation errors. Within using attention maps over the source/reference sentences for COMET-22 and UNITE, we find that the scores for the *src* only version are the worst suggesting that these metrics use very limited information from the source (c.f. the similar observation made in Section 7.2).

While using the length heuristic for the contrastive evaluation, GEMBA-MQM has better results followed by UNITE. As GEMBA-MQM and XCOMET-XL also provide labels to their predicted error spans, we also convert these labels into score based on the weights in Guerreiro et al. (2023) (critical: 10, major: 5, minor: 1), then cap the error score per sentence at 25, and finally convert the score to a value between 0 and 1. We find that weighted label scores have a good improvement over the length heuristic suggesting that more sophisticated heuristics need to be developed in the future to obtain better meta-evaluation strategies. After using the label weighted score, we find that the performance for XCOMET-XL is still lower than the performance in Table 6, suggesting that the scores produced by the joint model may not necessarily rely on the error spans produced by that model. In contrast, GEMBA-MQM improves on its performance in Tables 6 and 10. We attribute this to either a change in the underlying model powering GPT-4 between submissions to WMT and re-running for SPAN-ACES or the use of a different weighting scheme. We also find it encouraging, that *GEMBA-MQM* improves over *GEMBA-DA*, providing us with some evidence that label-based evaluation can be helpful.

We speculate that these poor results may be attributed to (i) the unavailability of labelled MQM data during training (COMET-22 and UNITE), (ii) the availability of labelled data for only

	COMET-22		UNITE				XCOMET-XL		GEMBA-MQM	
	src-ref	ref	src	src-ref	ref	src	length	weight	length	weight
Span Extraction Evaluation										
Span F1	26.9	26.2	4	22.7	22.7	7.3	10.6	10.6	8.67	8.67
Contrastive Evaluation										
addition	0.598	0.477	-0.177	0.522	0.475	0.317	-0.269	-0.191	-0.077	0.103
mistranslation	-0.313	-0.364	-0.482	-0.447	-0.431	-0.308	-0.222	-0.016	0.005	0.240
real-world knowledge	-0.470	-0.501	-0.417	-0.360	-0.377	-0.279	-0.202	0.088	-0.330	0.328
untranslated	-0.641	-0.056	-0.689	-0.759	0.260	-0.910	-0.239	-0.166	-0.152	0.103
do not translate	0.500	0.340	-0.380	0.460	0.520	0.380	0.060	0.100	-0.080	0.140
undertranslation	-0.192	-0.206	-0.392	0.110	0.092	-0.220	-0.066	0.250	0.162	0.368
overtranslation	-0.144	-0.174	-0.362	0.312	0.284	-0.088	0.008	0.430	0.236	0.554
omission	-0.770	-0.842	-0.838	-0.814	-0.784	-0.700	-0.381	-0.197	0.165	0.385
punctuation	-0.385	-0.479	-0.609	-0.642	-0.574	-0.624	-0.593	-0.525	0.039	0.129
wrong language	0.406	0.289	-0.212	0.484	0.387	0.285	-0.225	-0.279	-0.132	-0.047
ACES-Score	-4.3	-5.5	-13.0	-1.8	-1.1	-5.6	-5.3	1.1	1.8	8.8

Table 10

Results of span-based metrics on SPAN-ACES for the tasks of span extraction and then contrastive evaluation on ACES using the predicted spans as outlined in Section 6.4.1. Under COMET-22 and UniTE, use of src and ref denotes if these components were used to obtain attention weights which were converted to spans. Span-F1 is only calculated for the incorrect translation. For the contrastive evaluation on ACES, all the above methods consider a candidate translation to be better than the other translation if the number of predicted spans in the former translation is less than the later, denoted by "length". For the "weight" version of XCOMET-XL and GEMBA-MQM, the labels denoting error severity of the predicted spans are converted to a weighted score. We note the derived metrics - COMET-22 and UNITE have better results on the span extraction task than the metrics designed to predict the spans. This trend flips for the contrastive evaluation. Overall, all of the methods struggle on both tasks.

a few language pairs (XCOMET-XL), (iii) the use of proprietary models, and thus no knowledge of underlying training data (GEMBA-MQM), (iv) the fact that these metrics are the earliest designs for span-based evaluation, and (v) the fact that our annotation schemes and evaluation regimes are also the first of their kind, potentially introducing new challenges for span-based evaluation metrics. We also caution the readers that our heuristics for contrastive evaluation only offer a starting point. Future work can include model confidence, different weighting schemes, POS tags *etc.,* to compare the two translations.

7. Analysis

Aside from high-level evaluations of which metrics perform best, we are mostly interested in weaknesses of metrics in general that we can identify using ACES. This section shows an analysis of some general questions that we aim to answer using ACES.

7.1 How sensitive are the metrics to error types?

One important quality of a reliable metric is its ability to assign sufficiently different scores to a good vs. an incorrect translation. To evaluate and compare the difference between the scores that the metrics assign to the good and incorrect translations, we normalise the metric scores to a common scale with an open-ended range, using the statistics from the metric scores submitted to the 2022 and 2023 editions of the WMT metrics task (Freitag et al. 2022, 2023). We do that by scaling the metric scores based on the mean and IQR (Interquartile range) of the scores of that metric submitted to the WMT22/23 metric shared task (see Equation 3).

Examples	addition	omission	mistranslation	untranslated	do not translate	overtranslation	undertranslation	real-world knowledge	wrong language	punctuation
	931	951	22530	1187	76	962	967	2924	1840	
BLEURT-20	0.106	0.355	0.200	1.743	0.398	0.142	-0.002	0.055	**0.826**	0.318
COMET-20	0.073	0.410	0.262	1.486	0.312	0.150	0.061	0.051	-0.322	0.229
YISI-1	0.118	0.293	0.075	**2.575**	0.294	-0.036	-0.049	-0.044	0.190	**0.376**
metricx_xl_DA_2019	0.100	0.447	0.496	1.772	0.559	0.689	0.366	0.498	0.752	0.302
metricx_xl_MQM_2020	-0.056	0.361	**0.651**	0.422	**0.697**	**1.000**	**0.654**	**0.740**	-0.560	0.331
metricx_xxl_MQM_2020	-0.008	0.294	0.550	0.649	0.688	0.826	0.485	0.629	-0.768	0.225
COMETKiwi	**0.126**	0.441	0.594	0.699	0.272	0.572	0.358	0.337	-0.559	0.247
Cross-QE	0.104	0.422	0.599	-0.285	0.055	0.703	0.456	0.225	-0.510	0.109
UniTE-src	0.096	**0.524**	0.484	-0.782	0.318	0.394	0.254	0.191	-0.552	0.196
COMETKiwi	0.196	0.618	0.721	-0.180	0.285	0.719	0.439	0.316	-0.767	0.218
MS-COMET-QE-22	-0.040	0.391	0.258	**2.730**	0.126	0.257	0.185	0.095	-0.654	0.226
BLEURT-20	0.094	0.314	0.177	1.545	0.353	0.126	-0.002	0.048	**0.732**	**0.281**
MetricX-23-c	0.021	0.413	0.645	0.334	0.399	0.593	0.330	0.766	-0.728	0.131
MetricX-23	0.001	0.184	0.407	-0.022	0.363	0.675	0.367	0.491	-0.618	0.151
XCOMET-Ensemble	0.070	0.342	0.434	0.208	0.249	0.462	0.308	0.358	-0.713	0.151
GEMBA-MQM	**0.584**	**1.132**	**1.566**	2.380	**0.719**	**1.646**	**0.976**	**1.814**	0.328	0.226
MetricX-23-QE	0.001	0.410	0.903	0.100	0.309	0.995	0.660	0.955	-1.163	0.135
CometKiwi-XXL	0.118	0.519	0.713	2.038	0.197	0.550	0.306	0.611	-0.733	0.187

Table 11

Metric sensitivity scores (scaled by WMT scores, then avg(good - bad)) for the nine top level categories in the ACES ontology, plus the additional fluency category: punctuation. The double horizontal line delimits the metrics submitted to WMT 2022 (top three groups) and the metrics submitted to WMT 2023 (bottom three groups). In each of these groups, the horizontal lines delimit baseline metrics (top), participating reference-based metrics (middle) and participating reference-free metrics (bottom), where we picked the top three metrics from each. The highest result for each category is denoted by bold text with a green highlight.

$$score^* = \frac{score - Avg(score_{wmt})}{IQR_{wmt}} \qquad (3)$$

We calculate the sensitivity score of the metric (see Equation 4) on a subset of the ACES samples, which is the average difference between the normalised scores assigned to good translations and to incorrect translations. Here $score^*_+$ is the normalised score assigned to the *good translation* and $score^*_-$ is the normalised score assigned to the *incorrect translation*. The value range of the sensitivity scores is open. [16]

$$sensitivity = Avg(score^*_+ - score^*_-) \qquad (4)$$

Similarly to the Kendall's tau-like correlation scores, we then report the average score overall examples belonging to each of the nine top-level accuracy categories in ACES, plus the fluency category *punctuation*, calculated for the top three metrics from the baseline, reference-based and reference-free metrics each, submitted to WMT 2022, and WMT 2023 (see Table 11). The phenomena-level sensitivity scores for all the metrics submitted to WMT 2022 and WMT 2023 can be found in Appendix 9.

The average sensitivity scores of the metrics support the results reached by the analysis of the average Kendall's tau-like correlation scores in most cases. One of the most significant exceptions to that is that GEMBA-MQM has significantly higher sensitivity scores across a majority of the high-level phenomena when evaluated according to the average sensitivity scores, unlike the Kendall's tau-like correlation results.

Looking at the average sensitivity scores of the metrics in the last row of Tables I.1 and I.2 in Appendix 9, we can see that the metrics are more sensitive to the untranslated category than all the other categories by a margin, where the untranslated category is not one of the easier categories according to the average Kendall's tau-like correlation scores.

Regarding the subcategories of mistranslation, discourse, previously considered the least challenging category based on Kendall's tau-like correlation, emerges as the most difficult for the metrics according to sensitivity scores. It can be seen that across multiple 2022 and 2023 metrics, the average sensitivity scores of the metrics on the hallucination subcategory are higher compared to the average sensitivity scores on discourse, while the average Kendall's tau-like correlation scores favour the discourse subcategory over hallucination.

7.2 How sensitive are metrics to the source?

We designed our challenge sets for the type of ambiguous translation in a way that the correct translation candidate given an ambiguous reference can only be identified through the source sentence. See the third example in Table 1, where the reference is in non-gendered language, thus requiring the information in the source sentence about the female baker to disambiguate the sentence. We present a targeted evaluation intended to provide some insights into how important the source is for different metrics. For brevity, we include top three performing metrics in each category in 2022 and 2023, and a couple of baseline metrics. Table 12 shows the detailed results of each metric on the considered phenomena.

The most important finding is that the reference-free metrics generally perform much better on these challenge sets than the reference-based metrics. This indicates that reference-based

---

16 Evaluation scripts are available here: https://github.com/EdinburghNLP/ACES

metrics rely too much on the reference. Interestingly, most of the metrics that seem to ignore the source do not randomly guess the correct translation (which is a valid alternative choice when the correct meaning is not identified via the source) but rather they strongly prefer one phenomenon over the other. For example, several metrics show a gender bias either towards female occupation names (female correlations are high, male low) or male occupation names (vice versa). Likewise, most metrics prefer translations with frequent senses for the word-sense disambiguation challenge sets, although the difference between frequent and infrequent is not as pronounced as for gender.

Only metrics that look at the source and exhibit fewer such preferences can perform well on average on this collection of challenge sets. XCOMET-ENSEMBLE performs best out of the reference-based metrics and XCOMET-QE-ENSEMBLE performs best of all reference-free metrics. It is noteworthy that there is still a considerable gap between these two models across most of the error categories, suggesting that reference-based models should pay more attention to the source when a reference is ambiguous to reach the performance of reference-free metrics.

This finding is also supported by our real-world knowledge commonsense challenge set. If we compare the scores on the examples where the subordinate clauses are missing from both the source and the reference to the ones where they are only missing from the reference, we can directly see the effect of disambiguation through the source. The corresponding correlation gains are shown in Table H.1 in the Appendix. All reference-based model correlation scores improve less than most reference-free correlations when access to the subordinate clause is given through the source. This highlights again that reference-based metrics do not give enough weight to the source sentence.

| | since | | female | | male | | wsd | | |
	causal	temp.	anti.	pro.	anti.	pro.	freq.	infreq.	AVG
Examples	*106*	*106*	*1000*	*806*	*806*	*1000*	*471*	*471*	*4766*
BERTScore	-0.434	0.434	-0.614	-0.216	0.208	0.618	0.214	-0.223	-0.001
COMET-22	-0.415	0.792	**0.940**	**1.000**	-0.628	0.374	**0.558**	0.040	0.333
MS-COMET-22	-0.604	0.623	0.296	0.640	-0.342	0.046	0.316	-0.155	0.102
UniTE	**0.038**	-0.075	-0.890	-0.213	0.377	0.934	0.270	-0.223	0.027
MetricX-23	-1.000	**1.000**	-0.864	-0.062	0.062	0.870	0.227	-0.222	0.001
MetricX-23-c	-0.849	0.849	-0.998	-0.581	**0.576**	**0.996**	0.150	-0.133	0.172
XCOMET-Ensemble	-0.585	0.981	0.852	0.948	0.273	0.922	0.554	**0.231**	**0.522**
Cross-QE	**0.208**	0.830	0.976	0.995	-0.337	0.364	**0.762**	0.355	0.519
MS-COMET-QE-22	-0.283	0.792	-0.194	0.320	0.246	0.694	0.465	0.002	0.255
UniTE-src	-0.321	0.906	0.976	0.980	0.171	0.736	0.622	0.346	0.552
CometKiwi	0.075	**1.000**	**0.990**	**0.998**	-0.171	0.440	0.740	0.384	0.557
KG-BERTScore	0.075	**1.000**	**0.990**	**0.998**	-0.171	0.440	0.702	0.460	0.315
MetricX-23-QE-b	-0.566	0.868	0.968	0.995	**0.722**	**0.968**	0.643	**0.490**	0.643
XCOMET-QE-Ensemble	-0.208	0.925	0.930	0.975	0.546	0.912	0.740	0.477	**0.662**

Table 12
Results on the challenge sets where the good translation can only be identified through the source sentence. Upper block: reference-based metrics, lower block: reference-free metrics. The best results for each phenomenon and each group of models are marked in bold and green and the average overall can be seen in the last column.

7.3 How much do metrics rely on surface overlap with the reference?

Another question we are interested in is whether neural reference-based metrics still rely on surface-level overlap with the reference. For this analysis, we use the dataset we created for hallucinated
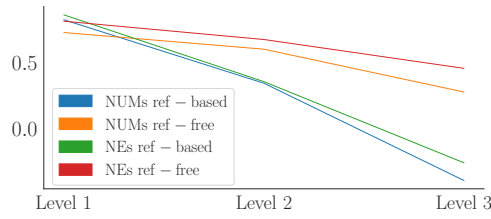
Figure 2
Decrease in correlation for reference-based and reference-free metrics on the named entity and number hallucination challenge sets.

named entities and numbers. We add an example about the three levels in Table 7.3 Note that as the levels increase, the surface level similarity between the good translation and the reference decreases while the surface level overlap between the incorrect translation and the reference increases.

SRC (es):	Sin embargo, Michael Jackson, Prince y **Madonna** fueron influencias para el álbum.
REF (en):	Michael Jackson, Prince and **Madonna** were, however, influences on the album.
Level-1 ✓:	However, Michael Jackson, Prince, and **Madonna** were influences on the album.
Level-1 ✗:	However, Michael Jackson, Prince, and **Garza** were influences on the album.
Level-2 ✓:	However, Michael Jackson, Prince, and **Madonna** were influences on the album.
Level-2 ✗:	Michael Jackson, Prince and **Garza** were, however, influences on the album.
Level-3 ✓:	The record was influenced by **Madonna**, Prince, and Michael Jackson though.
Level-3 ✗:	Michael Jackson, Prince and **Garza** were, however, influences on the album.

We take the average correlation for all reference-based metrics, (excluding lexical overlap metrics like BLEU) and the average correlation of all reference-free metrics that cover all languages across both the years and plot the decrease in correlation with increasing surface-level similarity of the incorrect translation to the reference. The result can be seen in Figure 2.

We can see that on average reference-based metrics have a much steeper decrease in correlation than the reference-free metrics as the two translation candidates become more and more lexically diverse and the surface overlap between the incorrect translation and the reference increases. This indicates a possible weakness of reference-based metrics: If one translation is lexically similar to the reference but contains a grave error while others are correct but share less surface-level overlap with the reference, the incorrect translation may still be preferred.

We also show that this is the case for the challenge set where we use an adversarial paraphrase from PAWS-X that shares a high degree of lexical overlap with the reference but does not have the same meaning as an incorrect translation. On average, the reference-based metrics only reach a correlation of 0.05 ± 0.17 on this challenge set, whereas the reference-free metrics reach a correlation of 0.24 ± 0.17. This shows that reference-based metrics are less robust when the incorrect translation has high lexical overlap with the reference.

	reference-based	reference-free
hallucination	-0.21 ± 0.15	+0.01 ± 0.05
overly-literal	-0.32 ± 0.16	+0.07± 0.09
untranslated	-0.43 ± 0.15	-0.00 ± 0.07

Table 13
Average correlation difference and standard deviation between the challenge sets with reference-copied good translations and the challenge sets with the synonymous good translations.
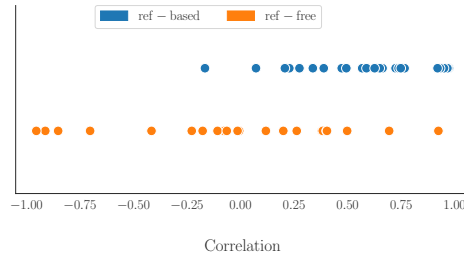
Figure 3
Correlation of reference-based metrics (blue) and reference-free metrics (orange) on the sentence-level untranslated test challenge set

.

7.4 Do multilingual embeddings help design better metrics?

As the community moves towards building metrics that use multilingual encoders, we investigate if some (un)desirable properties of multilingual embeddings or other base models are propagated in these metrics.

Multilingual models often learn cross-lingual representations by abstracting away from language-specific information (Wu and Dredze 2019). We are interested in whether the representations are still language-dependent in neural MT evaluation metrics which are trained on such models. For this analysis, we look at the sentence-level untranslated text challenge set (see Figure 3) and wrong language phenomena (see Table 5).

Figure 3 shows the correlations for all reference-based and reference-free metrics. Unsurprisingly, some reference-free metrics struggle considerably on this challenge set and almost always prefer the copied source to the real translation. The representations of the source and the incorrect translation are identical, leading to a higher surface and embedding similarity, and thus a higher score. We do, however, find some exceptions to this trend - COMET-KIWI and MS-COMET-QE-22 both have a high correlation on sentence-level untranslated text. This suggests that these metrics could have learnt language-dependent representations.

Most reference-based metrics have good to almost perfect correlation and can identify the copied source quite easily. As reference-based metrics tend to ignore the source (see Section 7.3), the scores are based on the similarity between the reference and the MT output. In this challenge set, the similarity between the good translation and the reference is likely to be higher than the incorrect translation and the reference. The former MT output is in the same language as the reference and will have more surface-level overlap. We believe the reference here acts as grounding.

However, this grounding property of the reference is only robust when the source and reference languages are dissimilar, as is the case with language pairs in the sentence-level untranslated text
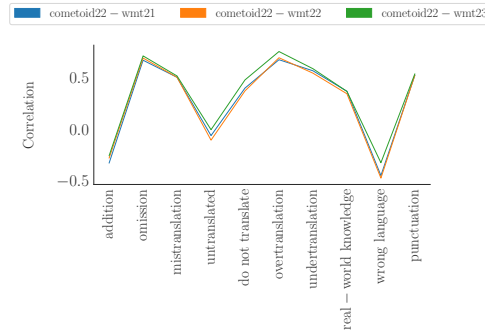
Figure 4
Correlations for different top-level phenomena categories with different models trained on successively more data.

challenge set. We find that reference-based metrics struggle on wrong language phenomena (see Tables 5, 8) where the setup is similar, but now the incorrect translation and the reference are from similar languages (e.g. one is in Hindi and the other is in Marathi). Naturally, there will be surface-level overlap between the reference and both the good translation and the incorrect translation. For example, both Marathi and Hindi use named entities with identical surface form, and so these will appear in the reference and also in both the good translation and the incorrect translation. Thus, the semantic content drives the similarity scores between the MT outputs and the references. The human translation in the similar language (labelled as the incorrect translation) may have a closer representation to the human reference because in the MT output (labelled as the good translation) some semantic information may be lost. We leave further investigation of this for future work.

7.5 How does metric training data size affect MT evaluation?

The COMETOID22 submission in 2023 included three different reference-free metric versions, each trained on successively more data. This allows us to investigate the effects of the metric training data size[17] on the performance on ACES. (Note that we cannot draw any conclusions about the training data size of the pretraining models that are used.) In Figure 4, we can see the effect of training data size on the performance on the top-level phenomena categories. COMETOID22-WMT23, the model that has seen the most data outperforms the other two metrics on almost all top-level categories. The correlation gain is especially pronounced for the *untranslated*, *do not translate* (content in the source is erroneously translated into the target language), *overtranslation* (the target translation contains more specific information than the source) and *wrong language* categories (see Table 1 for examples for each of the phenomena). For clearer insights as to where the performance gain comes from, we would need to analyse the training data in depth. However, it is evident from these results that more training data is beneficial for metric development. In the next section, we look at metric score changes over metric implementation cycles - where likely more than just the training data changed.

---

17 Note that for COMETOID22 this is not human judgement labelled data but rather pseudo labelled data where labels come from the reference-based COMET-22 model.

7.6 Changes between 2022 and 2023

We compare the results of metrics submitted by the same teams in both 2022 and 2023 in Table 14. To make a valid comparison, we exclude the examples affected by the double quote pre-processing resulting in 33817 examples which are discussed below.

	COMETKiwi		KG-BERTScore	XCOMET		
	-XL	-XXL		-Ensemble	-XL	-XXL
addition	-0.120	-0.004	-0.251	0.595	0.455	0.142
omission	-0.004	-0.002	0.103	0.118	-0.126	-0.254
mistranslation	-0.005	0.013	0.077	0.126	0.038	0.005
untranslated	0.000	0.142	0.266	-0.181	-0.342	-0.362
do not translate	-0.395	-0.553	0.000	0.053	0.079	-0.105
overtranslation	0.027	0.035	0.119	0.073	-0.067	0.017
undertranslation	-0.019	-0.021	0.077	0.014	-0.132	-0.025
real-world knowledge	-0.020	0.100	0.107	0.003	-0.123	-0.198
wrong language	-0.014	-0.173	-0.618	-0.296	-0.232	-0.395
punctuation	-0.037	0.004	0.264	0.206	-0.144	0.006
ACES-Score	-1.04	-0.38	0.40	4.23	0.21	-1.64

Table 14
Comparison of average Kendall's tau-like correlation: delta calculated as 2023 score minus 2022 score.

We report changes in performance in terms of deltas, computed by subtracting the 2022 score from the 2023 score. We do this for the following pairs of metrics: KG-BERTSCORE (2022) and KG-BERTSCORE (2023); COMETKIWI (2022) paired with COMETKIWI-XL (2023) and COMETKIWI-XXL (2023); COMET-22 (2022) paired with XCOMET-ENSEMBLE (2023), XCOMET-XL (2023) and XCOMET-XXL (2023).

We observe that the performance of KG-BERTSCORE improved in 2023. From the description provided by the metric developers, the main difference is that the 2023 version of KG-BERTSCORE metric uses COMET-QE instead of BERTScore (Zhang et al. 2020) to compute the similarity between the source and the hypothesis. Whilst we might therefore attribute the increase in performance to this change, a more systematic comparison of the two metric versions would be required to confirm whether this is the only contributing factor.

The metrics in the COMETKIWI family exhibit: a slight drop in performance (COMETKIWI-XL) and a similar performance to that of last year (COMETKIWI-XXL). The difference can be attributed to changing the underlying encoder, XLM-R XL and XLM-R XXL (Goyal et al. 2021) respectively, and the use of additional fine-tuning data made available this year. We have seen that the addition of more training data helps in Section 7.5. Considering that there is no improvement in the performance, we question if an increase in the underlying model capacity of the encoder alone is useful for obtaining better MT evaluation.

Performance change for the XCOMET family is variable: there is a performance increase for XCOMET-ENSEMBLE (compared to COMET-22), for XCOMET-XL the increase is smaller, and the performance of XCOMET-XXL is degraded. The XCOMET family is designed to provide both a quality score and an error span. Considering that the metric also provides an explanation of the scores without hurting the performance, this is indeed a positive change. Finally, it is worth noting that for *all* metrics in Table 14 a change in performance is observed for almost all ACES categories, for all metrics.

Whilst it is not possible to draw conclusions or make predictions about the future of metric development based solely on the observations from two consecutive metrics shared tasks, we

highlight several high-level changes. Firstly, we note the participation of many more COMET-based metrics in 2023, compared with 2022. This is presumably based on the success of COMET at previous shared tasks and its adoption within the MT community. We find that three metrics from 2022 were used as baseline metrics in 2023: COMET-22, COMETKIWI, and MS-COMET-QE-22. In contrast to the submissions in 2022, the 2023 submissions included some new metrics that use lexical overlap through text matching or embeddings (TOKENGRAM_F, PARTOKENGRAM_F, and EBLEU). However, their performance trend is similar to other surface overlap metrics. This year has also seen submissions based on large language models (EMBED_LLAMA and GEMBA-MQM). As seen in Section 6.1, their moderate performance indicates the need for more effective approaches. Additionally, we note an overall increase between 2022 and 2023 in the number of metrics submitted to WMT that a) provide segment-level scores and b) provide scores for all language pairs and directions in ACES. There were 37 segment-level metrics at WMT 2022, 24 of which covered the language pairs and directions in ACES, compared with 47 and 33, respectively in 2023. This suggests that the interest in metric development remains high, and could be increasing. In terms of metric sensitivities to error types, computed by subtracting the 2022 score from the 2023 score mostly support the observations about the changes in the metric performances evaluated using the Kendall's tau-like correlation.

In our overview analysis in Section 6.1 we highlight many similarities between metric performance trends in 2022 and 2023. However, there are a few differences. In 2023 we observe that the reference-free metric group performs strongly overall, compared with the baseline and reference-free groups. This could be attributed to the increase in metrics based on the COMET architecture in 2023. WMT 2023 also saw the submission of two LLM-based metrics: EMBED_LLAMA and GEMBA-MQM. Despite the success of LLMs across various tasks (Brown et al. 2020), the performance of both EMBED_LLAMA and GEMBA-MQM highlights that leveraging LLMs to evaluate translated outputs on segment-level still requires some improved design strategies. All of these observations suggest that evaluating MT outputs is indeed a hard problem (Neubig 2022). While we do have a good suite of metrics to provide a proxy for evaluation, there are indeed several interesting challenges that need to be tackled before we find an ideal evaluation regime. And even then, we need to continuously monitor this to ensure that we do not optimise towards metric weaknesses that we have not yet discovered.

8. Recommendations

Based on the metrics results on ACES, SPAN-ACES and our analyses, we first make some recommendations for MT evaluation in general and then provide some more specific suggestions for metric development.

Informative Evaluation: From our results in Section 6, we find that *a single score is not enough* to identify if a metric has superior performance. By evaluating on ACES, we can obtain a profile for the metric showcasing its strengths and weaknesses across different MT errors, supporting metric developers in making more informed choices. To further deter the development of metrics that produce a single score, we also recommend predicting error spans (ideally with labels) instead of scores. We propose SPAN-ACES as an additional test suite for the development of metrics that produce error spans.

Building metric ensembles: Both the evaluation on phenomena and language pair categories in Section 6 showed that there is no single best-performing metric. This divergence is likely to become even larger if we evaluate metrics on different domains. In future work on MT evaluation, it may be worthwhile thinking about how different metrics can be *combined* to make more robust decisions as to which is the best translation. Recent submissions to the WMT Metrics shared task include ensemble models (such as) COMET-KIWI, KG-BERTSCORE, XCOMET-ENSEMBLE, *etc.,*), which suggests that our recommendations are aligned with the efforts of the community.

The source matters: Our analysis in Section 7.2 highlighted that many reference-based metrics that take the source as input do not consider it enough. Cases where the correct translation can only be identified through the source, are currently better handled by reference-free metrics. This is a serious shortcoming of reference-based metrics which should be addressed in future research, also considering that many reference-based metrics choose to exclude source information by design.

Surface overlap prevails: In Section 7.3, we showed that despite moving beyond a purely surface-level comparison with the reference, most reference-based metrics are still considerably influenced by surface-level overlap. We thus recommend including paraphrases in the training regime as well as designing loss functions that explicitly discourage surface-level overlap.

Check the base model properties: Some properties of multilingual representations, like the representation space being language-agnostic, can result in undesirable effects on MT evaluation (Section 7.4). We also find that LLMs are not effective segment-level MT evaluators just yet (see Section 5.1), hence, better design strategies must be employed to make LLMs useful in evaluation. Simple strategies to model language-specific information in the metrics could also improve the robustness of the metrics to adversarial language pair attacks.

9. Conclusion

In this work, we identify and address some of the shortcomings of MT metrics. A single segment-level (or system-level) score for a metric does not provide an overview of that metric's strengths and weaknesses. To address this, we developed ACES: a translation accuracy challenge set based on the MQM ontology, which consists of 36,476 examples covering 146 language pairs and representing challenges from 68 phenomena. ACES can be used to provide a profile of metric performance over a range of phenomena, and to measure incremental performance between multiple versions of the same metric. We used ACES to evaluate the baseline and submitted metrics from the WMT 2022 and 2023 metrics shared tasks, to measure the incremental performance of those metrics submitted in both years, to measure how sensitive metrics are to certain phenomena, and to provide fine-grained analyses of metric performance to reveal the extent to which metrics rely on the source and on surface-level overlap with the reference, and to assess whether multilingual embeddings are a helpful component in metric design.

Our overview of metric performance at the phenomena and language levels in Section 6 reveals that there is no single best-performing metric. The more fine-grained analyses in Section 7 highlight that 1) metric sensitivity is correlated with score prediction for most of the metrics 2) many reference-based metrics that take the source as input do not consider it enough, 3) most reference-based metric scores are still considerably influenced by surface overlap with the reference, 4) the use of multilingual embeddings can have undesirable effects on MT evaluation and 5) the addition of metric-specific data improves the quality of the metric. We find that LLM-based evaluation methods have mediocre results and in some cases even worse than the surface overlap-based metrics.

We recommend that these shortcomings of existing metrics be addressed in future research and that metric developers should consider a) combining metrics with different strengths, e.g. in the form of ensemble models, b) developing metrics that give more weight to the source and less to surface-level overlap with the reference, and c) incorporating strategies to explicitly model additional language-specific information (rather than simply relying on multilingual embeddings). We also recommend the community develop evaluation methods that produce error types and error spans as singular scores are not informative. To that end, we released SPAN-ACES where every incorrect translation in ACES contains span-level annotations for the erroneous text corresponding to the phenomenon label. We also provided baseline results on SPAN-ACES. We have made

ACES and SPAN-ACES publicly available and hope that it will provide a useful benchmark for MT evaluation metric developers in the future.

Appendix A: Language Codes

Code	Language	Code	Language	Code	Language	Code	Language
af	Afrikaans	fa	Persian	ja	Japanese	sl	Slovenian
ar	Arabic	fi	Finnish	ko	Korean	sr	Serbian
be	Belarusian	fr	French	lt	Lithuanian	sv	Swedish
bg	Bulgarian	ga	Irish	lv	Latvian	sw	Swahili
ca	Catalan	gl	Galician	mr	Marathi	ta	Tamil
cs	Czech	he	Hebrew	nl	Dutch	th	Thai
da	Danish	hi	Hindi	no	Norwegian	tr	Turkish
de	German	hr	Croatian	pl	Polish	uk	Ukranian
el	Greek	hu	Hungarian	pt	Portuguese	ur	Urdu
en	English	hy	Armenian	ro	Romanian	vi	Vietnamese
es	Spanish	id	Indonesian	ru	Russian	wo	Wolof
et	Estonian	it	Italian	sk	Slovak	zh	Chinese

Table A.1
ISO 2-Letter language codes of the languages included in the challenge set

Appendix B: Permitted Unit Conversions

The unit conversions permitted for the *Hallucination - Unit Conversion* challenge set are listed in Table B.1.

Appendix C: Distribution of Examples Across Language Pairs

Table C.1 contains the total number of examples per language pair in the challenge set. As can be seen in the table, the distribution of examples is variable across language pairs. The dominant language pairs are: en-de, de-en, and fr-en.

Appendix D: Distribution of Language Pairs Across Phenomena

Table D.1 contains the list of language pairs per phenomena in the challenge set. As can be seen in the table, the distribution of language pairs is variable across phenomena. Addition and omission have the highest variety of language pairs. en-de is the most frequent language pair across all phenomena.

Appendix E: ACES Annotation Methods per Phenomena

The methods used to annotate the error spans for each of the phenomena in SPAN-ACES are listed in Table E.1.

Distance:	**Volume**:
• miles → metres	• barrels → gallons
• kilometres → miles	• barrels → litres
• kilometres → metres	• gallons → barrels
• metres → feet	• gallons → litres
• metres → yards	
• feet → metres	**Weight**:
• feet → yards	• kilograms → grams
• centimetres → inches	• kilograms → pounds
• centimetres → millimetres	• grams → ounces
• inches → centimetres	• ounces → grams
• inches → millimetres	
• millimetres → centimetres	**Time**:
• millimetres → inches	• hours → minutes
	• minutes → seconds
Speed:	• seconds → minutes
• miles per hour → kilometres per hour	• days → hours
• kilometres per hour → miles per hour	• months → weeks
• kilometres per second → miles per second	• weeks → days
• miles per second → kilometres per second	
Area:	
• square kilometres → square miles	

Table B.1
Permitted Unit Conversions

Appendix F: ACES Span Annotation Guidelines

1. General guidelines

Your task is to annotate spans of translation errors that match a specific error type: e.g. "word swap", or "overtranslation". You are presented with two sentences (A and B) as well as a label denoting the error type that you should look for. You should compare translations A and B and mark any error spans of the specified type that occur in sentence B.

Please note that:

- You should annotate at the word level, not at the character level. I.e. in the case that the error is a misspelling (e.g. "combuter" instead of "computer") the complete word ("combuter") should be marked.
- You should *only* mark errors of the type specified by the error type label, and no other errors that may be present in sentence B.
- You are *not* required to mark any errors that may be present in sentence A.
- Whilst the majority of sentences you will encounter will be fluent, some machine-generated sentences will contain disfluencies.
- In the examples in this document, errors are highlighted in bold text to help make the examples clearer. You do *not* need to bold the error spans in your annotations.
- This document is intended to be comprehensive and cover the cases assigned across multiple annotators. As such, a batch that is assigned to you may contain only a subset of the error types listed in the *Error type-specific* section (below).

Table C.1
Number of examples per language pair. Rows: source language; Columns: target language

src \ tgt	af	ar	be	bg	ca	cs	da	de	el	en	es	et	fa	fi	fr	ga	gl	he	hi	hr	hu	hy	id	it	ja	ko	lt	lv	mr	nl	no	pl	pt	ro	ru	sk	sl	sr	sv	sw	ta	th	tr	uk	ur	vi	wo	zh
af										96				25																																		
ar										361					102																																	
be										67																																						
bg										393	175								17																													
ca										79																	40																					
cs										85																																						
da										83	84				394										113	63									104													75
de										4163																																						
el										387																																						
en	25	5	6	15	347	368	46	6964	21		1263	25	20	12	800		16	18	343	27	44	3	31	10	430	545	17	19	52	50	53	349	44	46	698	27	45	15	39		1		10	16	10	25		333
es					88			64		1263					125										117	74																						67
et	16									70																																						
fa										85																																						
fi										79																																						
fr								683		2868	78														403	59			344						46													61
ga										17																																						
gl										70																																						
he										59																													51									
hi		8								367																																						
hr										81																																						
hu										53																		29																				
hy										48																																				13		
id										63																																						
it										801										24																												
ja								60		912	67				122											163																						74
ko								70		1004	72				110										358																							73
lt				28						68																																						
lv										61																																						
mr										63																																						
nl										73								28																														
no										53																						17																
pl										65																										58												
pt										89					42														111									40										
ro										91	87																						54															
ru								106		472																																						
sk										54																																						
sl										69																																						
sr										64																																						
sv										79																																						
sw										327																																						
ta										39																																						
th										299																																						
tr										386																																						
uk										77																																						
ur										372																																						
vi										391																																						
wo										11																																						
zh								150		1209	59				113										128	80																						

phenomena	language pairs	phenomena	language pair
ambiguous-translation-wrong-discourse-connective-since-causal	fr-en, de-en	hallucination-real-data-vs-ref-word	en-mr, de-en, en-de, fr-de
ambiguous-translation-wrong-discourse-connective-since-temporal		hallucination-real-data-vs-ref-synonym	en-de, de-en, fr-de
hallucination-unit-conversion-unit-matches-ref		untranslated-vs-ref-word	en-de, de-en, fr-de
ambiguous-translation-wrong-discourse-connective-while-contrast		untranslated-vs-synonym	en-de, de-en, fr-de
ambiguous-translation-wrong-discourse-connective-while-temporal		modal_verb:deletion	de-en
ambiguous-translation-wrong-gender-female-anti		modal_verb:substitution	de-en
ambiguous-translation-wrong-gender-male-anti		nonsense	ko-en, ko-ja, en-ko, fr-ja, de-en
ambiguous-translation-wrong-gender-male-pro		ordering-mismatch	en-de, de-en, fr-de
ambiguous-translation-wrong-sense-frequent	en-de, en-ru	overly-literal-vs-correct-idiom	en-de, de-en
ambiguous-translation-wrong-sense-infrequent	en-de, en-ru	overly-literal-vs-explanation	en-de, de-en
anaphoric_group_it-they:deletion	en-de	overly-literal-vs-ref-word	en-de, de-en, fr-de
anaphoric_group_it-they:substitution	en-de	overly-literal-vs-synonym	en-mr, de-en, en-de, fr-de
anaphoric_intra_non-subject_it:deletion	en-de	pleonastic_it:deletion	de-en
anaphoric_intra_non-subject_it:substitution	en-de	pleonastic_it:substitution_pro_trans_different_to_ref	de-en
anaphoric_intra_subject_it:deletion	en-de	punctuation:deletion_all	de-en
anaphoric_intra_subject_it:substitution	en-de	punctuation:deletion_commas	de-en
anaphoric_intra_they:deletion	en-de	punctuation:deletion_quotes	de-en
anaphoric_intra_they:substitution	en-de	punctuation:statement-to-question	de-en
anaphoric_singular_they:deletion	en-de	do-not-translate	
anaphoric_singular_they:substitution	en-de	real-world-knowledge-entailment	en-de, de-en
antonym-replacement	fr-en, ko-en, ja-en, es-en, zh-en, de-en	real-world-knowledge-hypernym-vs-distractor	en-de, de-en
similar-language-high	en-hi, en-cs, en-es	real-world-knowledge-hypernym-vs-hyponym	en-de, de-en
similar-language-low	fr-mr, en-pl, en-ca	real-world-knowledge-synonym-vs-antonym	en-de, de-en
coreference-based-on-commonsense	en-de, en-ru, en-fr	undertranslation	fr-en, vi-en, sw-en, tr-en, zh-en, ru-en, bg-en, el-en, th-en, es-en, hi-en, de-en, ar-en, ur-en
hallucination-named-entity-level-1	en-de, ja-de, en-ko, de-zh, ja-en, es-de, fr-en, es-ko, ko-ja, es-ja, de-ja, zh-es, fr-zh, fr-ja, es-en, fr-ko, zh-en, ko-de, ko-es, de-ko, ko-en, fr-es, ja-es, ja-ko, zh-fr, en-es, de-en, ja-fr, ko-zh, en-fr, de-fr, ko-fr, es-fr, zh-ko, fr-de, ja-zh, de-es, es-zh, en-ja, zh-de, en-zh, zh-ja	overtranslation	
hallucination-named-entity-level-2		xnli-addition-contradiction	en-de, vi-en, ca-es, en-et, hr-lv, da-en, no-en, uk-en, fi-en, en-da, ta-en, pl-en, gl-en, en-hr, hy-en, ur-en, fr-en, hr-en, lt-en, sr-sv, tr-en, en-no, en-sl, he-en, pl-sk, ru-en, ro-en, sv-en, en-lt, es-en, en-nl, nl-en, bg-en, he-sv, zh-en, hu-en, be-en, lv-hr, lv-en, bg-lt, en-ro, sk-pl, ko-en, ga-en, sk-en, af-en, sl-en, en-hu, sr-en, en-es, ca-en, en-sk, de-en, mr-en, id-en, vi-en, gl-en, en-fr, de-fr, pt-en, fr-de, en-pt, fa-en, hi-en, el-en, ar-en, it-en, en-pl, cs-en
hallucination-named-entity-level-3		xnli-addition-neutral	
hallucination-number-level-1		xnli-omission-contradiction	
hallucination-number-level-2		xnli-omission-neutral	
hallucination-number-level-3			
lexical-overlap	fr-en, en-fr, de-fr, ko-en, es-ja, ja-en, ko-fr, es-fr, ko-ja, de-ja, zh-en, ja-fr, zh-fr, en-ja, cs-en, fr-ja, de-en, zh-ja	hallucination-date-time	en-ur, en-hr, ur-en, en-no, en-sl, ro-en, en-vi, en-lt, es-en, en-nl, he-sv, en-it, en-ro, af-fa, en-id, lt-bg, en-af, af-en, es-ca, vi-en, sv-he, de-fr, pt-en, en-pl, et-en, hr-lv, wo-en, da-en, en-ko, en-da, ja-en, hy-en, pt-sr, hy-vi, fi-en, en-es, lt-en, en-sv, tr-en, bg-en, lv-en, bg-lt, sr-en, en-es, en-bg, en-pt, hi-en, el-en, it-en
hallucination-unit-conversion-amount-matches-ref	et-en, wo-en, da-en, no-en, uk-en, ta-en, fi-en, pl-en, ja-en, hy-en, ur-en, hr-en, fr-en, lt-en, tr-en, he-en, bg-en, ro-en, sv-en, ru-en, es-en, nl-en, zh-en, hu-en, be-en, lv-en, ko-en, ga-en, sk-en, af-en, sl-en, sr-en, ca-en, de-en, mr-en, id-en, vi-en, gl-en, pt-en, fa-en, hi-en, el-en, ar-en, it-en, cs-en		
hallucination-unit-conversion-unit-matches-ref		copy-source	ar-fr, ru-es, ur-en, fr-en, tr-en, zh-de, bg-en, ru-en, es-en, zh-en, sw-en, ja-ko, th-en, de-en, pl-mr, vi-en, hi-en, el-en, ar-en
commonsense-only-ref-ambiguous	en-de, fr-en, ru-fr, en-fr, de-fr, ru-de, fr-de, ru-en, en-ru, fr-ru, de-ru, de-en	addition	en-ur, en-hr, ur-en, en-no, en-sl, ro-en, en-vi, en-lt, es-en, en-nl, he-sv, en-it, en-ro, af-fa, en-id, lt-bg, en-af, af-en, es-ca, vi-en, sv-he, de-fr, pt-en, en-pl, et-en, hr-lv, wo-en, da-en, en-ko, en-da, ja-en, hy-en, pt-sr, hy-vi, fi-en, en-es, lt-en, en-sv, tr-en, bg-en, lv-en, bg-lt, sr-en, en-es, en-bg, en-pt, hi-en, el-en, it-en
commonsense-src-and-ref-ambiguous		omission	
addition	en-ca, en-el, en-et, en-ta, pl-en, hr-en, he-en, pl-sk, en-ar, ru-en, en-fi, zh-en, hu-en, be-en, lv-hr, en-he, ko-en, en-fa, sl-en, ca-en, en-gl, en-tr, en-sk, de-en, en-sr, fa-af, fa-en, ar-en, cs-en, en-de, en-hy, ar-hi, no-en, uk-en, fi-en, en-be, sr-pt, en-ru, sv-en, nl-en, sk-pl, en-hi, en-hu, mr-en, hi-ar, id-en, gl-en, en-fr, en-lv, fr-de, ca-es, en-uk,		
omission			

Table D.1

Collection of list of languages per phenomena

Phenomenon	Annotation Method
addition	addition/omissions
ambiguous-translation-wrong-discourse-connective-since-causal	word-lvl-compare-to-good
ambiguous-translation-wrong-discourse-connective-since-temporal	word-lvl-compare-to-good
ambiguous-translation-wrong-discourse-connective-while-contrast	word-lvl-compare-to-good
ambiguous-translation-wrong-discourse-connective-while-temporal	word-lvl-compare-to-good
ambiguous-translation-wrong-gender-female-anti	word-lvl-compare-to-good
ambiguous-translation-wrong-gender-female-pro	word-lvl-compare-to-good
ambiguous-translation-wrong-gender-male-anti	word-lvl-compare-to-good
ambiguous-translation-wrong-gender-male-pro	word-lvl-compare-to-good
ambiguous-translation-wrong-sense-frequent	word-lvl-compare-to-good
ambiguous-translation-wrong-sense-infrequent	word-lvl-compare-to-good
anaphoric_group_it-they:deletion	addition/omissions
anaphoric_group_it-they:substitution	addition/omissions
anaphoric_intra_non-subject_it:deletion	addition/omissions
anaphoric_intra_non-subject_it:substitution	addition/omissions
anaphoric_intra_subject_it:deletion	addition/omissions
anaphoric_intra_subject_it:substitution	addition/omissions
anaphoric_intra_they:deletion	addition/omissions
anaphoric_intra_they:substitution	addition/omissions
anaphoric_singular_they:deletion	addition/omissions
anaphoric_singular_they:substitution	addition/omissions
antonym-replacement	word-lvl-compare-to-ref
commonsense-only-ref-ambiguous	word-lvl-compare-to-good
commonsense-src-and-ref-ambiguous	word-lvl-compare-to-good
copy-source	whole-sentence
coreference-based-on-commonsense	manual
do-not-translate	word-lvl-compare-to-good
hallucination-date-time	date-time
hallucination-named-entity-level-1	word-lvl-compare-to-good
hallucination-named-entity-level-2	word-lvl-compare-to-ref
hallucination-named-entity-level-3	word-lvl-compare-to-ref
hallucination-number-level-1	word-lvl-compare-to-good
hallucination-number-level-2	word-lvl-compare-to-ref
hallucination-number-level-3	word-lvl-compare-to-ref
hallucination-real-data-vs-ref-word	manual
hallucination-real-data-vs-synonym	manual
hallucination-unit-conversion-amount-matches-ref	unit-conversion
hallucination-unit-conversion-unit-matches-ref	unit-conversion
hypernym-replacement	word-lvl-compare-to-ref
hyponym-replacement	word-lvl-compare-to-ref
lexical-overlap	manual
modal_verb:deletion	addition/omissions
modal_verb:substitution	word-lvl-compare-to-good
nonsense	word-lvl-compare-to-ref
omission	addition/omissions
ordering-mismatch	word-swap
overly-literal-vs-correct-idiom	word-lvl-compare-to-good
overly-literal-vs-explanation	word-lvl-compare-to-good
overly-literal-vs-ref-word	word-lvl-compare-to-good
overly-literal-vs-synonym	word-lvl-compare-to-good
pleonastic_it:deletion	addition/omissions
pleonastic_it:substitution	addition/omissions
punctuation:deletion_all	addition/omissions
punctuation:deletion_commas	addition/omissions
punctuation:deletion_quotes	addition/omissions
punctuation:statement-to-question	addition/omissions
real-world-knowledge-entailment	word-lvl-compare-to-good
real-world-knowledge-hypernym-vs-distractor	word-lvl-compare-to-good
real-world-knowledge-hypernym-vs-hyponym	word-lvl-compare-to-good
real-world-knowledge-synonym-vs-antonym	word-lvl-compare-to-good
similar-language-high	whole-sentence
similar-language-low	whole-sentence
untranslated-vs-ref-word	word-lvl-compare-to-good
untranslated-vs-synonym	word-lvl-compare-to-good
xnli-addition-contradiction	whole-sentence
xnli-addition-neutral	whole-sentence
xnli-omission-contradiction	whole-sentence
xnli-omission-neutral	whole-sentence

Table E.1
Methods used to annotate the error spans for each of the phenomena in SPAN-ACES

- You should only mark punctuation as part of error spans if it is part of the error (e.g. added as part of an addition operation or changed as part of a substitution operation).

Please read the guidelines thoroughly before you start the annotation task. Once you have finished, please make a second pass to identify and correct any mistakes that you may have made. Please also make a note of any examples that you were unsure how to annotate e.g. the example ID and a brief note.

All error spans should be marked with open and closing tags (e.g. <error span>). Errors of specific types may be formed by addition, substitution, deletion or reordering operations. For deletion operations, you should insert an empty pair of tags <> where content is missing in sentence B.

Whitespace: Error tags should *not* contain leading (e.g. **< error span>**) or trailing (e.g. **<error span >**) whitespace.

Addition: a text span that is not present in sentence A is included in sentence B.

Sentence A: The cat is a species of small carnivorous mammal.
Sentence B: The cat is a **<domestic>** species of small carnivorous mammal.

Substitution: a text span in sentence A is substituted with a different text span in sentence B.

Sentence A: Female domestic cats can have kittens from spring to late autumn.
Sentence B: Female domestic cats can have kittens from **<May>** to **<December>**.

Deletion: a text span that is present in sentence A is omitted from sentence B. Note that when marking a deletion, care should be taken to ensure that no extra whitespace is inserted into the sentence. Tags marking the deletion should be inserted after the space separating the two words where the deletion occurred.

Sentence A: Feral cats are domestic cats that were born in or have reverted to a wild state.
Sentence B: Feral cats are domestic cats **<>**or have reverted to a wild state.

Reordering: a text span in sentence A that appears in a different position in sentence B, as though the sentence has been reordered.

Sentence A: Montreal is the second most populous city in Canada and the most populous city in the province of Quebec.
Sentence B: Montreal is the **<>**most populous city in Canada and the **<second>** most populous city in the province of Quebec.

Note: reordering operations can be viewed as a combination of a *deletion* and an *addition* operation to change the order of elements of a sentence.

Example 1: Marking a single error span of a specified error type; ignoring other error types

In this example, the aim is to mark "overtranslation" type errors, i.e. where translation B is more specific than translation A:

Sentence A: The festival in Houston took place in the summer.
Sentence B: The festival in took place in August.

The error span is "August", which is more specific than "the summer" - the information that the event took place in August has been "hallucinated".

Annotated B: The Republican National Convention in was in **\<August\>**.

Note that the missing information in sentence B ("Houston") can be ignored because it is an "omission" error not an "overtranslation" error. Other examples of errors that can be ignored include e.g. agreement errors in German.

Example 2: Marking multiple error spans in the same example

If there are multiple errors of the specified type present in sentence B, you should mark each error span individually. For example, if the error label is "omission" you should mark the two spans of omitted text in sentence B:

Sentence A: Like the other planets in the Solar System, Mars was formed 4.5 billion years ago.
Sentence B: Like the other planets, Mars was formed 4.5 years ago.

Annotated B: Like the other planets **\<\>**, Mars was formed 4.5 **\<\>**years ago.

2. Error type-specific guidelines

In your annotations, you will only encounter three specific error types. Additional guidelines are provided below for these error types - hallucination, word swap and coreference.

Hallucination

In a *hallucination* example, text that is not present in sentence A is observed in sentence B or word in sentence A is replaced by a more frequent or orthographically similar word in sentence B. I.e. hallucination can be an "addition" or a "substitution" case. This may result in a change of meaning in sentence B. You should mark the "hallucinated" text in sentence B.

Sentence A: The official languages of Scotland are: English, Scots, and Scottish Gaelic.
Sentence B: The official languages of Scotland are: English, Welsh, French, Scots, and Scottish Garlic.

The information that Welsh and French are official languages of Scotland has been hallucinated and inserted into sentence B. Additionally, "Gaelic" has been hallucinated as "Garlic". This should be annotated as:

Annotated B: The official languages of Scotland are: English, **\<Welsh, French,\>** Scots, and Scottish **\<Garlic\>**.

Word Swap

In a *word swap* example the position of a word or a span of text in sentence A appears swapped in sentence B. This may result in sentence B being factually incorrect. You should mark (in sentence B) the spans of text that have been swapped.

> Sentence A: Their music is considered by many as an alternative metal with rap metal and industrial metal influences, which according to previous interviews call themselves "murder - rock".
> Sentence B: Their music is considered by many as industrial metal with rap metal and alternative metal influences. According to previous interviews, they consider themselves "murder rock".

The position of the words "alternative" and "industrial" is different in sentence A, compared with sentence B and should be annotated as follows:

> Annotated B: Their music is considered by many as **<industrial>** metal with rap metal and **<alternative>** metal influences. According to previous interviews, they consider themselves "murder rock".

Coreference

In a *coreference* example a pronoun in sentence A is replaced with a (potentially) inappropriate noun-phrase in sentence B. You should mark the relevant noun-phrase in sentence B.

Example:

> Sentence A: The cat had caught the mouse and it was trying to wriggle free.
> Sentence B: The cat had caught the mouse and the cat was trying to wriggle free.

The pronoun "it" has been replaced with the noun-phrase "the cat", resulting in a change in meaning. This should be annotated as:

> Annotated B: The cat had caught the mouse and **<the cat>** was trying to wriggle free.

Appendix G: Prompt for LLMs for MT evaluation

For reference-based evaluation, we used the following prompt:
Score the following translation with respect to human reference on a continuous scale of 0 to 100 where score of zero means "no meaning preserved" and score of one hundred means "perfect meaning and grammar". Only output an integer between 0 to 100.
Source: *source sentence here*
Human Reference: *reference sentence here*
Translation: *candidate translation*

For reference-free evaluation, we excluded the "with respect to human reference" and "Human Reference" from the prompt.

Appendix H: Importance of source

We report the results on the real-world knowledge commonsense challenge set in Table H.1. Reference-based metrics tend to disregard the information in the source.

Reference-based	corr-gain	Reference-free	corr-gain
BERTScore	0.002	COMET-QE	0.018
COMET-20	0.06	Cross-QE	0.292
COMET-22	0.19	HWTSC-Teacher-Sim	0.154
metricx_xxl_DA_2019	0.012	KG-BERTScore	0.154
metricx_xxl_MQM_2020	-0.016	MS-COMET-QE-22	0.196
MS-COMET-22	0.05	UniTE-src	0.216
UniTE	0.042	cometoid22-wmt23	0.138
COMET-22	0.042	CometKiwi	0.454
MetricX-23	0.004	CometKiwi-XL	0.148
MetricX-23-b	-0.002	GEMBA-MQM	1.107
MetricX-23-c	0.008	KG-BERTScore	0.436
XCOMET-Ensemble	0.162	MS-COMET-QE-22	0.198
XCOMET-XL	0.11	MetricX-23-QE-b	0.296
XCOMET-XXL	0.016	XCOMET-QE-Ensemble	0.112
		XLsimQE	0.184

Table H.1
Results on the real-world knowledge commonsense challenge set with reference-based metrics in the left block and reference-free metrics in the right block. The numbers are computed as the difference between the correlation with the subordinate clause in the source and the correlation without the subordinate clause in the source. Largest gains are bolded.

Examples	addition	omission	mistranslation	untranslated	do not translate	overtranslation	undertranslation	real-world knowledge	wrong language	punctuation
	931	*951*	*22530*	*1187*	*76*	*962*	*967*	*2924*	*1840*	*1449*
BLEU	**0.222**	0.253	-0.078	1.213	0.314	-1.093	-1.096	-0.293	0.655	0.365
f101spBLEU	0.136	0.186	-0.071	0.911	0.348	-0.563	-0.608	-0.160	0.503	0.211
f200spBLEU	0.131	0.181	-0.072	0.892	0.356	-0.529	-0.553	-0.159	0.496	0.192
chrF	0.061	0.253	-0.073	1.535	0.289	-0.286	-0.222	-0.087	0.656	0.107
BERTScore	0.115	0.182	-0.018	1.373	0.304	-0.018	-0.037	-0.026	0.375	**0.493**
BLEURT-20	0.106	0.355	0.200	1.743	0.398	0.142	-0.002	0.055	**0.826**	0.318
COMET-20	0.073	0.410	0.262	1.486	0.312	0.150	0.061	0.051	-0.322	0.229
COMET-QE	-0.103	0.117	0.072	-0.126	0.049	0.286	0.196	0.096	-0.310	0.026
YISI-1	0.118	0.293	0.075	**2.575**	0.294	-0.036	-0.049	-0.044	0.190	0.376
COMET-22	0.045	0.250	0.292	0.399	0.318	0.352	0.216	0.207	-0.497	0.217
metricx_xl_DA_2019	0.100	0.447	0.496	1.772	0.559	0.689	0.366	0.498	0.752	0.302
metricx_xl_MQM_2020	-0.056	0.361	**0.651**	0.422	**0.697**	**1.000**	**0.654**	**0.740**	-0.560	0.331
metricx_xxl_DA_2019	0.085	0.411	0.532	1.679	0.543	0.547	0.189	0.423	0.088	0.244
metricx_xxl_MQM_2020	-0.008	0.294	0.550	0.649	0.688	0.826	0.485	0.629	-0.768	0.225
MS-COMET-22	-0.048	0.351	0.133	0.987	0.183	0.222	0.119	0.059	-0.159	0.192
UniTE	0.113	0.420	0.321	0.534	0.353	0.338	0.152	0.189	-0.439	0.204
UniTE-ref	0.078	0.385	0.304	0.081	0.294	0.378	0.182	0.187	-0.436	0.168
COMETKiwi	0.126	0.441	0.594	0.699	0.272	0.572	0.358	0.337	-0.559	0.247
Cross-QE	0.104	0.422	0.599	-0.285	0.055	0.703	0.456	0.225	-0.510	0.109
HWTSC-Teacher-Sim	-0.005	0.140	0.163	-0.574	0.200	0.224	0.174	0.055	-0.063	0.239
HWTSC-TLM	-0.095	0.148	0.120	0.104	-0.019	0.293	0.241	0.062	-0.158	0.463
KG-BERTScore	0.139	0.236	0.428	-0.786	0.598	0.186	0.124	0.079	-0.170	0.366
MS-COMET-QE-22	-0.038	0.369	0.243	2.564	0.119	0.241	0.174	0.090	-0.616	0.212
UniTE-src	0.096	**0.524**	0.484	-0.782	0.318	0.394	0.254	0.191	-0.552	0.196
Average	0.062	0.309	0.259	0.794	0.327	0.209	0.076	0.142	-0.066	0.251

Table I.1

Metric sensitivity scores (scaled by WMT scores, then avg(good - bad)) of metrics submitted to WMT 2022 for the nine top level categories in the ACES ontology, plus the additional fluency category: punctuation. The horizontal lines delimit baseline metrics (top), participating reference-based metrics (middle) and participating reference-free metrics (bottom). The best result for each category is denoted by bold text with a green highlight. Note that *Average* is an average over averages.

Examples	addition	omission	mistranslation	untranslated	do not translate	overtranslation	undertranslation	real-world knowledge	wrong language	punctuation
	931	951	22530	1187	76	962	967	2924	1840	1449
BERTScore	0.119	0.189	-0.010	1.410	0.311	-0.047	-0.066	-0.031	0.384	**0.505**
BLEU	0.131	0.149	-0.046	0.716	0.185	-0.646	-0.647	-0.173	0.387	0.216
BLEURT-20	0.094	0.314	0.177	1.545	0.353	0.126	-0.002	0.048	0.732	0.281
chrF	0.053	0.220	-0.064	1.338	0.252	-0.249	-0.194	-0.076	0.572	0.094
COMET-22	0.036	0.320	0.199	0.846	0.272	0.250	0.127	0.107	-0.213	0.196
COMETKiwi	0.196	0.618	0.721	-0.180	0.285	0.719	0.439	0.316	-0.767	0.218
f200spBLEU	0.121	0.167	-0.066	0.824	0.329	-0.489	-0.511	-0.147	0.458	0.177
MS-COMET-QE-22	-0.040	0.391	0.258	**2.730**	0.126	0.257	0.185	0.095	-0.654	0.226
Random-sysname	-0.003	0.016	0.003	0.013	0.015	0.023	-0.004	-0.008	0.003	-0.013
YISI-1	0.114	0.283	0.072	2.489	0.284	-0.034	-0.048	-0.043	0.183	0.365
eBLEU	0.070	0.135	0.014	1.358	0.169	-0.199	-0.194	-0.068	**0.986**	0.065
embed_llama	0.190	0.324	0.027	0.962	0.262	-0.217	-0.851	-0.360	0.139	0.130
MetricX-23	0.001	0.184	0.407	-0.022	0.363	0.675	0.367	0.491	-0.618	0.151
MetricX-23-b	-0.029	0.231	0.375	0.135	0.385	0.601	0.336	0.460	-0.696	0.140
MetricX-23-c	0.021	0.413	0.645	0.334	0.399	0.593	0.330	0.766	-0.728	0.131
tokengram_F	0.054	0.214	-0.058	1.335	0.280	-0.281	-0.231	-0.080	0.603	0.208
XCOMET-Ensemble	0.070	0.342	0.434	0.208	0.249	0.462	0.308	0.358	-0.713	0.151
XCOMET-XL	0.047	0.244	0.303	-0.125	0.227	0.310	0.198	0.288	-0.650	0.075
XCOMET-XXL	0.057	0.260	0.349	-0.197	0.210	0.476	0.375	0.298	-0.690	0.119
XLsim	0.085	0.220	0.126	1.442	0.372	-0.129	-0.158	-0.063	0.192	0.365
cometoid22-wmt21	-0.063	0.233	0.207	-0.513	0.138	0.399	0.260	0.147	-0.528	0.215
cometoid22-wmt22	-0.058	0.243	0.221	-0.567	0.134	0.396	0.260	0.148	-0.545	0.205
cometoid22-wmt23	-0.044	0.270	0.239	-0.479	0.155	0.304	0.198	0.106	-0.600	0.198
CometKiwi-XL	0.094	0.581	0.703	2.047	0.264	0.459	0.284	0.486	-0.614	0.197
CometKiwi-XXL	0.118	0.519	0.713	2.038	0.197	0.550	0.306	0.611	-0.733	0.187
GEMBA-MQM	**0.584**	**1.132**	**1.566**	2.380	**0.719**	**1.646**	**0.976**	**1.814**	0.328	0.226
KG-BERTScore	0.175	0.550	0.638	-0.181	0.341	0.639	0.388	0.277	-0.685	0.173
MetricX-23-QE	0.001	0.410	0.903	0.100	0.309	0.995	0.660	0.955	-1.163	0.135
MetricX-23-QE-b	0.001	0.441	0.836	0.145	0.269	0.764	0.527	0.930	-1.123	0.124
MetricX-23-QE-c	-0.011	0.291	0.609	-0.072	0.169	0.695	0.485	0.963	-0.759	0.098
XCOMET-QE-Ensemble	0.080	0.373	0.517	0.322	0.196	0.491	0.335	0.389	-0.734	0.112
Average	0.073	0.332	0.355	0.722	0.265	0.308	0.143	0.290	-0.266	0.183

Table I.2

Metric sensitivity scores (scaled by WMT scores, then avg(good - bad)) of metrics submitted to WMT 2023 for the nine top level categories in the ACES ontology, plus the additional fluency category: punctuation. The horizontal lines delimit baseline metrics (top), participating reference-based metrics (middle) and participating reference-free metrics (bottom). The best result for each category is denoted by bold text with a green highlight. Note that *Average* is an average over averages.

Appendix I: Phenomena-level Metric Sensitivity Scores

Tables I.1 and I.2 contain the average sensitivity scores for each high-level phenomena of the metrics submitted to WMT 2022 and WMT 2023 respectively.

References

Alves, Duarte, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

Amrhein, Chantal, 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, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

Amrhein, Chantal, Nikita Moghe, and Liane Guillou. 2023. Aces: Translation accuracy challenge sets at wmt 2023. In *Proceedings of the Eighth Conference on Machine Translation*, pages 693–710, Association for Computational Linguistics, Singapore.

Amrhein, Chantal and Rico Sennrich. 2022. Identifying weaknesses in machine translation metrics through minimum bayes risk decoding: A case study for COMET. In *2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, Association for Computational Linguistics, Online.

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

Avramidis, Eleftherios, Vivien Macketanz, Arle Lommel, and Hans Uszkoreit. 2018. Fine-grained evaluation of quality estimation for machine translation based on a linguistically motivated test suite. In *Proceedings of the AMTA 2018 Workshop on Translation Quality Estimation and Automatic Post-Editing*, pages 243–248, Association for Machine Translation in the Americas, Boston, MA.

Avramidis, Eleftherios, Shushen Manakhimova, Vivien Macketanz, and Sebastian Möller. 2023. Challenging the state-of-the-art machine translation metrics from a linguistic perspective. In *Proceedings of the Eighth Conference on Machine Translation*, pages 713–729, Association for Computational Linguistics, Singapore.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Association for Computational Linguistics, Austin, Texas.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Association for Computational Linguistics, Berlin, Germany.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Campolungo, Niccolò, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352, Association for Computational Linguistics, Dublin, Ireland.

Carlini, Nicholas, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. In

USENIX Security Symposium.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Iacer Calixto, John Tinsley, and Andy Way. 2017. Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108:109–120.

Chen, Xiaoyu, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Ting Zhu, Mengli Zhu, Ning Xie, Lizhi Lei, Shimin Tao, Hao Yang, and Ying Qin. 2022. Exploring robustness of machine translation metrics: A study of twenty-two automatic metrics in the WMT22 metric task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 530–540, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

Chia, Yew Ken, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models.

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Association for Computational Linguistics, Brussels, Belgium.

Dale, David, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50, Association for Computational Linguistics, Toronto, Canada.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota.

Dréano, Sören, Derek Molloy, and Noel Murphy. 2023a. Embed_llama: Using llm embeddings for the metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 736–743, Association for Computational Linguistics, Singapore.

Dréano, Sören, Derek Molloy, and Noel Murphy. 2023b. Tokengram_f, a fast and accurate token-based chrf++ derivative. In *Proceedings of the Eighth Conference on Machine Translation*, pages 728–735, Association for Computational Linguistics, Singapore.

Dziri, Nouha, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaïd Harchaoui, and Yejin Choi. 2023. Faith and fate: Limits of transformers on compositionality. *ArXiv*, abs/2305.18654.

ElNokrashy, Muhammad and Tom Kocmi. 2023. ebleu: Unexpectedly good machine translation evaluation using simple word embeddings. In *Proceedings of the Eighth Conference on Machine Translation*, pages 744–748, Association for Computational Linguistics, Singapore.

Emelin, Denis and Rico Sennrich. 2021. Wino-X: Multilingual Winograd schemas for commonsense reasoning and coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8517–8532, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.

Fan, Angela, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

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

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang

Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Freitag, Markus, 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, Association for Computational Linguistics, Singapore.

Freitag, Markus, 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, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

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

Gowda, Thamme, Tom Kocmi, and Marcin Junczys-Dowmunt. 2023. Cometoid: Distilling strong reference-based machine translation metrics into even stronger quality estimation metrics. In *Proceedings of the Eighth Conference on Machine Translation*, pages 749–753, Association for Computational Linguistics, Singapore.

Goyal, Naman, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. pages 29–33.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Guerreiro, Nuno M., Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection.

Guillou, Liane and Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 636–643, European Language Resources Association (ELRA), Portorož, Slovenia.

Guillou, Liane, 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, Association for Computational Linguistics, Belgium, Brussels.

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

Isabelle, Pierre, 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, Association for Computational Linguistics, Copenhagen, Denmark.

Jia, Robin and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Association for Computational Linguistics, Copenhagen, Denmark.

Juraska, Juraj, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. Metricx-23: The google submission to the wmt 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 754–765, Association for Computational Linguistics, Singapore.

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

Khashabi, Daniel, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 252–262, Association for Computational Linguistics, New Orleans, Louisiana.

King, Margaret and Kirsten Falkedal. 1990. Using test suites in evaluation of machine translation systems. In *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

Kocmi, Tom and Christian Federmann. 2023a. Gemba-mqm: Detecting translation quality error spans with gpt-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 766–773, Association for Computational Linguistics, Singapore.

Kocmi, Tom and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, European Association for Machine Translation, Tampere, Finland.

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

Kocmi, Tom, Hitokazu Matsushita, and Christian Federmann. 2022. MS-COMET: More and Better Human Judgements Improve Metric Performance. In *Proceedings of the Seventh Conference on Machine Translation*, Association for Computational Linguistics, Abu Dhabi.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Koehn, Philipp and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, Association for Computational Linguistics, New York City.

Kudo, Taku and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Association for Computational Linguistics, Brussels, Belgium.

Laali, Majid and Leila Kosseim. 2017. Improving discourse relation projection to build discourse annotated corpora. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 407–416, INCOMA Ltd., Varna, Bulgaria.

Lapshinova-Koltunski, Ekaterina, Christian Hardmeier, and Pauline Krielke. 2018. ParCorFull: a parallel corpus annotated with full coreference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan.

Li, Yitong, Trevor Cohn, and Timothy Baldwin. 2017. BIBI system description: Building with CNNs and breaking with deep reinforcement learning. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 27–32, Association for Computational Linguistics, Copenhagen, Denmark.

Liu, Yilun, Xiaosong Qiao, Zhanglin Wu, Su Chang, Min Zhang, Yanqing Zhao, shimin tao Song Peng, Hao Yang, Ying Qin, Jiaxin Guo, Minghan Wang, Yinglu Li, Peng Li, and Xiaofeng Zhao. 2022. Partial Could Be Better Than Whole: HW-TSC 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Association for Computational Linguistics, Abu Dhabi.

Lo, Chi-kiu. 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, Association for Computational Linguistics, Florence, Italy.

Lo, Chi-kiu, 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 774–797, Association for Computational Linguistics, Singapore.

Lommel, Arle, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, 0:455–463.

Lu, Qingyu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt.

Mahler, Taylor, Willy Cheung, Micha Elsner, David King, Marie-Catherine de Marneffe,

Cory Shain, Symon Stevens-Guille, and Michael White. 2017. Breaking NLP: Using morphosyntax, semantics, pragmatics and world knowledge to fool sentiment analysis systems. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 33–39, Association for Computational Linguistics, Copenhagen, Denmark.

McCoy, Richard T and Tal Linzen. 2019. Non-entailed subsequences as a challenge for natural language inference. *Proceedings of the Society for Computation in Linguistics (SCiL)*, pages 358–360.

Moghe, Nikita, Tom Sherborne, Mark Steedman, and Alexandra Birch. 2023. Extrinsic evaluation of machine translation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13060–13078, Association for Computational Linguistics, Toronto, Canada.

Mukherjee, Ananya and Manish Shrivastava. 2022. REUSE: REference-free UnSupervised quality Estimation Metric. In *Proceedings of the Seventh Conference on Machine Translation*, Association for Computational Linguistics, Abu Dhabi.

Mukherjee, Ananya and Manish Shrivastava. 2023. Mee4 and xlsim : Iiit hyd's submissions' for wmt23 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 798–803, Association for Computational Linguistics, Singapore.

Neubig, Graham. 2022. Is my nlp model working? the answer is harder than you think.

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

Papineni, Kishore, 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, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA.

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

Perrella, Stefano, Lorenzo Proietti, Alessandro Scirè, Niccolò Campolungo, and Roberto Navigli. 2022b. Machine Translation Evaluation as a Sequence Tagging Problem. In *Proceedings of the Seventh Conference on Machine Translation*, Association for Computational Linguistics, Abu Dhabi.

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

Popović, Maja and Sheila Castilho. 2019. Challenge test sets for MT evaluation. In *Proceedings of Machine Translation Summit XVII: Tutorial Abstracts*, European Association for Machine Translation, Dublin, Ireland.

Raganato, Alessandro, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 470–480, Association for Computational Linguistics, Florence, Italy.

Ravichander, Abhilasha, Siddharth Dalmia, Maria Ryskina, Florian Metze, Eduard Hovy, and Alan W Black. 2021. NoiseQA: Challenge set evaluation for user-centric question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2976–2992, Association for Computational Linguistics, Online.

Rei, Ricardo, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. 2023. The inside story: Towards better understanding of machine translation neural evaluation metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105, Association for Computational Linguistics, Toronto, Canada.

Rei, Ricardo, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 Submission for the Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*,

Association for Computational Linguistics, Abu Dhabi.

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

Rimell, Laura, Stephen Clark, and Mark Steedman. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 813–821, Association for Computational Linguistics, Singapore.

Rios, Annette, 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, Association for Computational Linguistics, Belgium, Brussels.

Rocchietti, Guido, Flavia Achena, Giuseppe Marziano, Sara Salaris, and Alessandro Lenci. 2021. Fancy: A diagnostic data-set for nli models. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it)*.

Rudinger, Rachel, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Association for Computational Linguistics, Valencia, Spain.

Scao, Teven Le, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Sellam, Thibault, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. 2020.

Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 921–927, Association for Computational Linguistics, Online.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Association for Computational Linguistics, Berlin, Germany.

Sinha, Koustuv, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.

Smith, Noah A. 2012. Adversarial evaluation for models of natural language. *CoRR*, abs/1207.0245.

Specia, Lucia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91, Association for Computational Linguistics, Online.

Staliūnaitė, Ieva and Ben Bonfil. 2017. Breaking sentiment analysis of movie reviews. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 61–64, Association for Computational Linguistics, Copenhagen, Denmark.

Stanovsky, Gabriel, 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, Association for Computational Linguistics, Florence, Italy.

Tao, Shimin, Su Chang, Ma Miaomiao, Hao Yang, Xiang Geng, Shujian Huang, Min Zhang, Jiaxin Guo, Minghan Wang, and Yinglu Li. 2022. CrossQE: HW-TSC 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 646–652, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid).

Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin,

Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Toral, Antonio and Víctor M. Sánchez-Cartagena. 2017. A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1063–1073, Association for Computational Linguistics, Valencia, Spain.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Vamvas, Jannis and Rico Sennrich. 2021. Contrastive conditioning for assessing disambiguation in MT: A case study of distilled bias. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10265, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.

Vamvas, Jannis and Rico Sennrich. 2022. As little as possible, as much as necessary: Detecting over- and undertranslations with contrastive conditioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 490–500, Association for Computational Linguistics, Dublin, Ireland.

Vieira, Lucas Nunes, Minako O'Hagan, and Carol O'Sullivan. 2021. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases.

Information, Communication & Society, 24(11):1515–1532.

Wan, Yu, Keqin Bao, Dayiheng Liu, Baosong Yang, Derek F. Wong, Lidia S. Chao, Wenqiang Lei, and Jun Xie. 2022a. Alibaba-Translate China's Submission for WMT2022 Metrics Shared Task. In *Proceedings of the Seventh Conference on Machine Translation*, Association for Computational Linguistics, Abu Dhabi.

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

Wu, Shijie and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Association for Computational Linguistics, Hong Kong, China.

Wu, Zhanglin, Yilun Liu, Min Zhang, Xiaofeng Zhao, Junhao Zhu, Ming Zhu, Xiaosong Qiao, Jingfei Zhang, Ma Miaomiao, Zhao Yanqing, Song Peng, shimin tao, Hao Yang, and Yanfei Jiang. 2023. Empowering a metric with llm-assisted named entity annotation: Hw-tsc's submission to the wmt23 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 820–826, Association for Computational Linguistics, Singapore.

Xu, Wenda, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5967–5994, Association for Computational Linguistics, Singapore.

Yang, Yinfei, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Association for Computational Linguistics, Hong Kong, China.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore:

Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, Association for Computational Linguistics, New Orleans, Louisiana.

Zhou, Jianing, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Association for Computational Linguistics, Online.