XCOMET: Transparent Machine Translation Evaluation through Fine-grained Error Detection

Nuno M. Guerreiro[*1,3,4,5], **Ricardo Rei**[*1,2,5], **Daan van Stigt**[1], **Luisa Coheur**[2,5], **Pierre Colombo**[4], **André F. T. Martins**[1,3,5]

[1]Unbabel Lisbon Portugal [2]INESC-ID, Lisbon, Portugal
[3]Instituto de Telecomunicações, Lisbon, Portugal
[4]MICS, CentraleSupélec, Université Paris-Saclay, France
[5]Instituto Superior Técnico, University of Lisbon, Portugal

Abstract

Widely used learned metrics for machine translation evaluation, such as COMET and BLEURT, estimate the quality of a translation hypothesis by providing a single sentence-level score. As such, they offer little insight into translation errors (e.g., what are the errors and what is their severity). On the other hand, generative large language models (LLMs) are amplifying the adoption of more granular strategies to evaluation, attempting to detail and categorize translation errors. In this work, we introduce **XCOMET**, an open-source learned metric designed to bridge the gap between these approaches. **XCOMET** integrates both sentence-level evaluation and error span detection capabilities, exhibiting state-of-the-art performance across all types of evaluation (sentence-level, system-level, and error span detection). Moreover, it does so while highlighting and categorizing error spans, thus enriching the quality assessment. We also provide a robustness analysis with stress tests, and show that XCOMET is largely capable of identifying localized critical errors and hallucinations.

1 Introduction

Automatic metrics for machine translation evaluation are widely used by researchers and practitioners to evaluate the quality of translations and the systems generating them. Notably, *learned* neural metrics, such as COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020), have demonstrated significant improvements in terms of correlation with human judgments when compared to traditional metrics like BLEU (Papineni et al., 2002; Freitag et al., 2021b, 2022).

These metrics are trained to regress on scores obtained through human annotations, by predicting a single sentence-level score representing the quality of the translation hypothesis. However, these single scores do not offer a detailed view into translation errors (e.g., it is not immediate which words or spans of words are wrongly translated). Moreover, as they are obtained by making use of highly complex pre-trained models, they can be difficult to interpret (Rei et al., 2023b; Leiter et al., 2023). One appealing strategy to bring a more detailed view into translation errors is to obtain finer-grained information on error spans through highlighting them and indicating their severity (Fonseca et al., 2019; Perrella et al., 2022; Bao et al., 2023). In fact, this is the strategy adopted in recent work that has employed generative large language models (LLMs) for machine translation evaluation: (i) identify errors within a given translation, subsequently (ii) categorize these errors according to their severity, and finally (iii) infer a sentence-level score from the predicted errors (Fernandes et al., 2023; Xu et al., 2023). However, these methods still lag behind dedicated learned metrics when using open LLMs, such as the LLaMA models (Touvron et al., 2023; Xu et al., 2023). As it stands, competitive performance with generative strategies remains contingent on utilizing large *proprietary*, *closed* LLMs such as PaLM-2 and GPT-4 (Fernandes et al., 2023; Kocmi and Federmann, 2023a).

In this work, we bridge the gap between these two approaches to machine translation evaluation by introducing **XCOMET**: a *learned* metric that simultaneously performs sentence-level evaluation and error span detection. Through extensive experiments, we show that our metrics leverage the strengths of both paradigms: They achieve state-of-the-art performance in all relevant vectors

[*]Equal contribution. Corresponding authors:
✉ nuno.guerreiro, ricardo.rei@unbabel.com.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 979–995, 2024. https://doi.org/10.1162/tacl_a_00683
Action Editor: Colin Cherry. Submission batch: 12/2023; Revision batch: 4/2024; Published 9/2024.

of evaluation (sentence-level, system-level, and error span prediction), while offering, via the predicted error spans, a lens through which we can analyze translation errors and better interpret the sentence-level scores. We achieve this by employing a curriculum during training that is focused on leveraging high-quality *publicly available* data at both the sentence- and error span level, complemented by synthetic data constructed to enhance the metric's robustness. Moreover, XCOMET is a unified metric (Wan et al., 2022b), supporting all modes of evaluation within a single model. This enables the metric to be used for quality estimation (when no reference is available), or for reference-only evaluation, similarly to BLEURT (when a source is not provided). Crucially, XCOMET also provides sentence-level scores that are directly inferred from the predicted error spans, in the style of AUTOMQM (Fernandes et al., 2023) and INSTRUCTSCORE (Xu et al., 2023).

Our contributions can be summarized as follows:

1. We introduce **XCOMET**, a novel evaluation metric that leverages the advantages of regression-based metrics and error span detection to offer a more detailed view of translation errors.
2. We show that **XCOMET** is a state-of-the-art metric at all relevant vectors of evaluation—sentence-level, system-level, and error span prediction—generally outperforming widely used neural metrics and generative LLM-based machine translation evaluation.
3. We provide a comprehensive robustness analysis of **XCOMET**, showing that this new suite of metrics identifies the vast majority of localized critical errors and hallucinations.
4. We release two evaluation models: **XCOMET-XL**, with 3.5B parameters, and **XCOMET-XXL**, featuring 10.7B parameters.[1]

2 Background

Methodologies for Human Assessment of Translation Quality. Human evaluation of machine translation is primarily conducted through three distinct approaches: post-edits (PE), direct assessments (DA), and the Multidimensional Quality Metrics (MQM) framework.

In post-edits (PE), *professional translators* are tasked with ''fixing'' a given translation, making minimal edits to improve its quality. Using this edited translation—often termed *post-edit*—we can evaluate the machine translation output by quantifying the number of edits, thus gauging the initial translation's quality (Snover et al., 2006).

Direct assessments (DA) (Graham et al., 2013) are a simple and widely-used evaluation method. Annotators—*non-expert bilingual speakers* or *professional translators*—are asked to annotate each translation with a score ranging from 0 to 100 to reflect its adequacy and fluency, where a score of 100 corresponds to a perfect translation, and 0 corresponds to a completely inadequate one.

The Multidimensional Quality Metrics (MQM) framework (Lommel et al., 2014), on the other hand, offers a more comprehensive and systematic approach to MT evaluation. *Professional translators* highlight errors—typically in the form of error spans—within translations, attributing them severity ratings (e.g., *minor*, *major*, or *critical*) and categorical labels (e.g., *fluency*, *accuracy*). Figure 1 illustrates one such annotation. MQM annotations have gained prominence in recent years due to their capacity to offer detailed insights into translation errors, facilitating more fine-grained and accurate comparisons between translation systems (Freitag et al., 2021a). As such, the field of Automatic Evaluation of MT has increasingly favored comparisons using MQM annotations over traditional DA and PE methodologies (Freitag et al., 2021b, 2022; Zerva et al., 2022).

Automatic Metrics for Translation Evaluation. Conventional automatic metrics for machine translation (MT) evaluation rely on *lexical*-based approaches, where the evaluation score is computed through statistics related to lexical overlap between a machine translation and a reference translation. Despite evidence indicating that these lexical metrics (e.g., BLEU [Papineni et al., 2002] and CHRF [Popović, 2015]) do not consistently align with human judgments, particularly when these are obtained through the MQM framework (Freitag et al., 2021b, 2022), they remain very popular. In fact, BLEU remains the most widely employed evaluation metric in machine translation to this day (Marie et al., 2021). On the other hand, *neural* metrics (e.g., COMET [Rei et al., 2020] and BLEURT [Sellam et al., 2020]) that rely on complex neural networks to estimate the quality of MT

[1] https://github.com/Unbabel/COMET/.

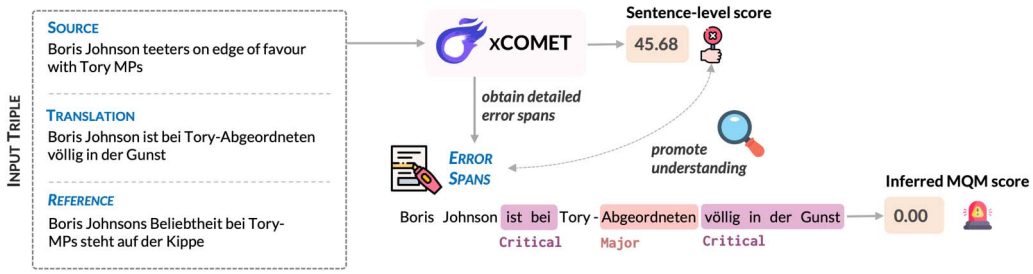


Figure 1: The **xCOMET** framework illustrated through a **real** example from the WMT22 News test set: The metric not only provides a sentence-level score, but also predicts translation error spans along with their respective severity. From these spans, we can infer MQM score (following the MQM typology), which informs and highly correlates with the sentence-level score (see Section 6). These spans complement the sentence-level score by providing a detailed view into the translation errors.

outputs are consistently among the best metrics for MT evaluation according to correlations with human judgments (Freitag et al., 2021b, 2022).

However, contrary to lexical metrics, which offer a straightforward interpretation, it can often prove challenging to explain the score predicted by a *neural* metric to a given translation output. As such, there have been a series of efforts to bring interpretability to neural metrics by focusing on understanding the inner workings of neural metrics (Rei et al., 2023b; Leiter et al., 2023), or on constructing inherently interpretable neural metrics (e.g., MaTESe [Perrella et al., 2022] and Fg-ted [Bao et al., 2023]) by assigning a central role to the task of predicting *word-level* errors in a given translation, instead of *just* a sentence-level score.

More recently, with the rise of generative LLMs, some studies have tried to frame the MT evaluation problem as a generative problem. This offers great flexibility, as the LLM can be prompted to either score the translation directly (Kocmi and Federmann, 2023b), or to identify errors in the translation (e.g., in line with the MQM framework) (Fernandes et al., 2023; Xu et al., 2023).

3 Problem Statement

An automatic *metric* for translation evaluation aims at predicting the quality of a translated sentence, $\boldsymbol{t}$, in light of a reference translation, $\boldsymbol{r}$, for a given source sentence, $\boldsymbol{s}$. Here, we focus specifically on neural metrics that make use of a neural model, and typically operate under one of the following evaluation scenarios:

- **reference-only** (ref): The model evaluates the translation by processing it alongside a ground-truth reference sentence (BLEURT is an example of such a metric);
- **source-reference combined input** (src+ref): The model evaluates the translation by jointly processing it with both the source and the reference (COMET is an example of such a metric);
- **source-only** (src): The model evaluates the translation using only its corresponding source sequence (COMETKIWI (Rei et al., 2022b) is an example of such a model). This mode is commonly termed as *quality estimation* (QE) or *reference-free* evaluation (Specia et al., 2010).

In essence, the model's input sequence consists of the translation $\boldsymbol{t}$ paired with some **additional input**—either $\boldsymbol{r}$, $[\boldsymbol{r}, \boldsymbol{s}]$, or $\boldsymbol{s}$—derived from the scenarios above. Given this input, the model may predict the quality of the translation at different granularities, e.g., sentence-level or word(span)-level.

Sentence-level Prediction. The model is tasked to predict a single global score (typically between 0 and 1) for the translation that represents how well it aligns with its context (i.e., source and/or reference sentence). These scores can be used for a broad range of tasks, such as gauging the quality of different translation systems (Freitag et al., 2022), identifying pathological translations (Guerreiro et al., 2023), assisting the generation of translations by MT systems (Fernandes et al., 2022), or even acting as reward models for human alignment of language models (Gulcehre et al., 2023).

Word(span)-level Prediction. In contrast, word-level (or span-level) predictions are more

fine-grained, identifying individual words or phrases in the translation that may have errors or discrepancies—typically identifying them as OK/BAD or according to their severity, e.g., MINOR/MAJOR. These granular evaluations are more interpretable and assist in pinpointing specific issues, which can be particularly valuable for feedback and iterative translation improvements.

Our metric, **XCOMET**, emerges in a unique position in the landscape of MT evaluation metrics. It can simultaneously perform evaluation under all three of the scenarios (SRC, REF, SRC+REF) presented, and provide sentence-level scores and error span annotations that are in line with the MQM framework, thus bringing further transparency to the evaluation (see Figure 1 for an illustration). In the next section, we detail the design choices and methodology of XCOMET.

4 Design and Methodology of XCOMET

In this section, we describe the methodology behind XCOMET, outlining its model architecture, training settings and corpora, and learning curriculum. We detail how the model is designed to perform both regression and error span detection while adopting a unified input approach for enhanced flexibility and performance.

4.1 Model Architecture

XCOMET is built upon insights from contributions to the WMT22 Metrics and QE shared tasks (Rei et al., 2022a,b). It is designed to concurrently handle two tasks: sentence-level regression and error span detection. Figure 2 illustrates its architecture. We follow the same architecture of the scaled-up version of COMETKIWI detailed in Rei et al. (2023a), which uses a large pre-trained encoder model as its backbone encoder model. Importantly, following from our multi-task setup, the model has two prediction heads: (i) a sentence-level *regression* head, which employs a feed-forward network to generate a sentence score, and (ii) a word-level *sequence tagger*, which applies a linear layer to assign labels to each translation token.

We train two **XCOMET** versions—**XCOMET-XL** and **XCOMET-XXL**—using the XL (3.5B parameters) and XXL (10.7B parameters) versions of XLM-R (Goyal et al., 2021).[2]

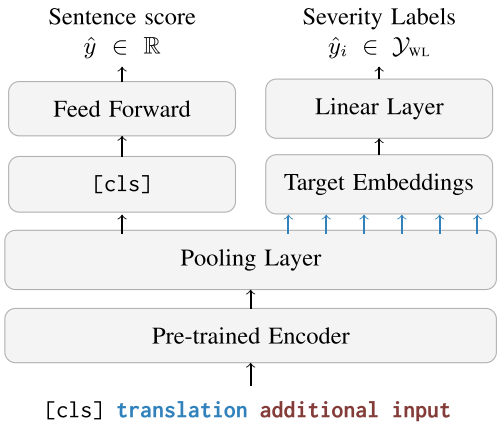


Figure 2: Architecture of **XCOMET**. The input to the model starts with a `[cls]` token followed by a **`translation`** and an **`additional input`** that will have the source, reference or both. After the pooling layer the `[cls]` token is passed to a feed-forward network to produce a quality score while all subword pieces corresponding to the **`translation`** are passed to a linear layer that will classify them according to their severity levels, $\mathcal{Y}_{\text{WL}} = \{\text{OK}, \text{MIN}, \text{MAJ}, \text{CRIT}\}$.

4.2 Fully Unified Evaluation

XCOMET adopts a unified input approach (Wan et al., 2022b), allowing for all the evaluation scenarios described in Section 3—REF, SRC+REF, and SRC evaluation—under a single model. Thus, the input sequence consists of two parts: (i) the translated sentence $\boldsymbol{t} = [t_1, \ldots, t_n]$ of length n, and (ii) an additional input containing information from the source, reference, or both. To do so, when a reference is available, we run three distinct forward passes (one for each evaluation scenario), each yielding sentence-level and word-level predictions.

4.2.1 Training Time

For each step, we collect the sentence-level predictions and the word-level logits for each input format: $\{\hat{y}_{\text{SL}}^{\text{SRC}}, \hat{y}_{\text{SL}}^{\text{REF}}, \hat{y}_{\text{SL}}^{\text{SRC+REF}}\}$ and $\{\hat{\boldsymbol{y}}_{\text{WL}}^{\text{SRC}}, \hat{\boldsymbol{y}}_{\text{WL}}^{\text{REF}}, \hat{\boldsymbol{y}}_{\text{WL}}^{\text{SRC+REF}}\}$.[3]

As we have mentioned before, XCOMET models are trained with supervision from both sentence-level quality assessments, y_{SL}, and word-level severity tags, $\boldsymbol{y}_{\text{WL}} = [y_1, \ldots, y_n]$, with $y_i \in \mathcal{Y}_{\text{WL}} = \{\text{OK}, \text{MIN}, \text{MAJ}, \text{CRIT}\}$. In the multi-task

[2]To the best of our knowledge, these represent the two largest open-source encoder-only models.

[3]Here, for each INPUT , $\in \{\text{SRC}, \text{REF}, \text{SRC+REF}\}$, we define $\hat{\boldsymbol{y}}_{\text{WL}}^{\text{INPUT}} = [\hat{y}_1^{\text{INPUT}}, \ldots, \hat{y}_n^{\text{INPUT}}]$.

setting, we use the following loss $\mathcal{L}$ for each input type (INPUT , $\in \{$SRC, REF, SRC+REF$\}$):

$$\mathcal{L}_{\text{SL}}^{\text{INPUT}} = (y_{\text{SL}} - \hat{y}_{\text{SL}}^{\text{INPUT}})^2 \quad (1)$$

$$\mathcal{L}_{\text{WL}}^{\text{INPUT}} = -\frac{1}{n}\sum_{i=1}^{n} \alpha_{y_i} \log p(y_i^{\text{INPUT}}) \quad (2)$$

$$\mathcal{L}^{\text{INPUT}} = (1-\lambda)\mathcal{L}_{\text{SL}}^{\text{INPUT}} + \lambda \mathcal{L}_{\text{WL}}^{\text{INPUT}} \quad (3)$$

$\boldsymbol{\alpha} \in \mathbb{R}^{|\mathcal{Y}_{\text{WL}}|}$ represents the class weights given for each severity label and λ is used to weigh the combination of the sentence and word-level losses.

The final learning objective is the summation of the losses for each input type:

$$\mathcal{L} = \mathcal{L}^{\text{SRC}} + \mathcal{L}^{\text{REF}} + \mathcal{L}^{\text{SRC+REF}} \quad (4)$$

Furthermore, in line with preceding metrics constructed upon the COMET framework, our models use features such as gradual unfreezing, and discriminative learning rates.

4.2.2 Inference Time

Error Span Prediction. For each subword in the translation, we average the output distribution of the word-level linear layer obtained for each forward pass. Using this distribution, we predict a set of word-level tags $\hat{\boldsymbol{y}}_{\text{WL}} = [\hat{y}_1, \ldots, \hat{y}_n]$ by taking the most likely class for each token. From these tags, we construct a list of *error spans*, S, by grouping adjacent subwords identified as errors. The severity of each span in S is defined according to the most severe error tag found within the span.

Sentence-level Prediction. For each forward pass, we obtain the corresponding sentence-level scores: $\hat{y}_{\text{SRC}}$, $\hat{y}_{\text{REF}}$, and $\hat{y}_{\text{SRC+REF}}$.[4] Additionally, we leverage the information coming from the predicted list of error spans, S, to infer an automated MQM score. To do so, we follow the MQM framework: we obtain the error counts for each severity level—$c_{\text{MIN}}, c_{\text{MAJ}}, c_{\text{CRIT}}$—and apply the predetermined severity penalty multipliers to define the error type penalty total, $e(S)$. Formally:

$$e(S) = c_{\text{MIN}} + 5 \times c_{\text{MAJ}} + 10 \times c_{\text{CRIT}} \quad (5)$$

Finally, we obtain $\hat{y}_{\text{MQM}}$ by capping and flipping the sign of $e(S)$:

$$\hat{y}_{\text{MQM}} = \begin{cases} \frac{25-e(S)}{25}, & \text{if } e(S) < 25 \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

[4]Here, for ease of notation, we use $\hat{y}_{\text{SRC}}$, $\hat{y}_{\text{REF}}$, and $\hat{y}_{\text{SRC+REF}}$ to represent the sentence-level score for each input type.

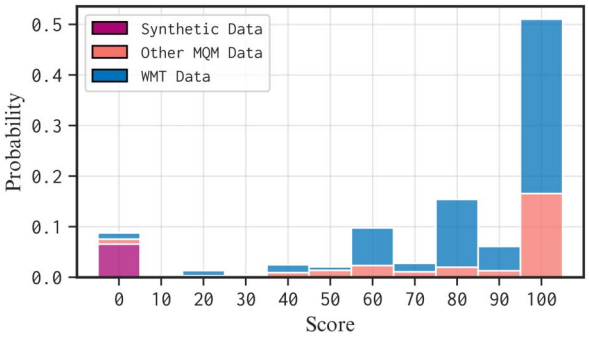


Figure 3: Histogram of the sentence-level scores from each partition of the MQM data. We aggregate the data from IndicMT, DEMETR under ''Other MQM Data''.

Note that the predicted score $\hat{y}_{\text{MQM}}$ is bounded between 0 and 1, with a score of 1 corresponding to a perfect translation.

We aggregate the scores to compute the final sentence-level score, $\hat{y}_{\text{SL}}$, through a weighted sum of the different sentence-level scores (see Figure 3). Importantly, we also include the inferred MQM score $\hat{y}_{\text{MQM}}$ to directly inform the final sentence-level prediction. Formally, given $\hat{\boldsymbol{y}} = [\hat{y}_{\text{SRC}}, \hat{y}_{\text{REF}}, \hat{y}_{\text{SRC+REF}}, \hat{y}_{\text{MQM}}]$:

$$\hat{y}_{\text{SL}} = \boldsymbol{w}^\top \hat{\boldsymbol{y}} \quad (7)$$

where $\boldsymbol{w}$ is set to $[1/9, 1/3, 1/3, 2/9]$.

4.3 Corpora

Our models are exclusively trained on publicly available DA and MQM annotations, most of which have been collected by WMT over the recent years.

DA Data. We use DA annotations collected by WMT from 2017 to 2020, and the MLQE-PE dataset (Fomicheva et al., 2022). As the MLQE-PE dataset does not contain reference translations, we used the post-edit translations as reference translations. Overall, the corpus consists of around 1 million samples, spanning 36 language pairs.

MQM Data. We collected the MQM annotations from WMT from 2020 to 2022.[5] We also used annotations sourced from other MQM-annotated datasets: (i) IndicMT (Sai B. et al., 2023), which contains MQM annotations spanning 5 Indian languages, and (ii) DEMETR (Karpinska et al., 2022), a diagnostic dataset with perturbations spanning semantic, syntactic, and morphological errors.

[5]Here, we exclude the 2022 News domain annotations, which we reserved for testing.

DATASET	No. Samples	Error Statistics
WMT Data	147K (76%)	63%; [57, 42, 1]
IndicMT	7K (4%)	80%; [19, 52, 29]
DEMETR	22K (11%)	47%; [38, 19, 43]
MLQE-PE Hall.	1.7K (1%)	All set to CRIT
Synthetic Hall.	16K (8%)	All set to CRIT

Table 1: Number of samples, as well as error statistics (overall percentage of non-correct translations; rates of error type [MIN, MAJ, CRIT]), of each MQM data source used for training XCOMET.

Corpora with MQM annotations are usually extremely unbalanced with critical errors being underrepresented (see stats for WMT in Table 1). As a result, metrics might struggle to effectively detect translations with critical errors and hallucinations. (Amrhein and Sennrich, 2022; Raunak et al., 2022; Guerreiro et al., 2023). As such, we augment the MQM corpus with hallucinations from the MLQE-PE corpus and *synthetic critical* errors. We create detached and oscillatory hallucinations (Raunak et al., 2021; Guerreiro et al., 2023): (i) detached hallucinations, replacing the translation with a random sentence or an unrelated one semantically similar to the source sentence;[6] and (ii) oscillatory hallucinations, where we randomly sample a n-gram from the translation (with n in $\{2, 3, 4\}$) and repeat it between 1 and 10 times. We set the sentence-level scores of these hallucinations to 0. Overall, our MQM corpus consists of 194K samples across 14 language pairs.

Scaling of Sentence-level Scores. While the sentence-level scores inferred from MQM annotations (through the procedure in Equation 6) are bounded between 0 and 1, DA annotations usually require z-normalization in order to mitigate variations in scoring strategies by different annotators (Bojar et al., 2017).[7] Thus, as z-scores are inherently centered at 0 and unbounded, there is a scaling mismatch between the data samples.

Consequently, to circumvent this limitation, we employ min-max scaling on our DA corpus to set its range of scores to $[0, 1]$. To do so, we set a practical minimum and maximum z-score value. We obtain the minimum score by averaging the z-scores for translations with over 1 annotation, wherein all annotators unanimously scored them with an unnormalized 0 DA score, i.e., they deemed the translation as ''random''. For determining a maximum value, we applied the same process for perfect translations, i.e., unnormalized 100 DA score.[8]

4.4 Training Curriculum

XCOMET models undergo a 3-phase curriculum training. Throughout these phases, the training emphasis alternates between sentence-level prediction and error span prediction by tweaking the parameter λ in Equation 3. The curriculum phases can be described as follows:

Phase I: The model is trained exclusively using the DA data. In this phase, the focus is exclusively set on sentence-level regression.

Phase II: In this stage, we introduce word-level supervision. To achieve this, the model is fine-tuned on our diverse MQM corpus, with most emphasis placed on the word-level task.

Phase III: The last training phase is aimed at unifying both tasks. The model is further fine-tuned using high-quality MQM data from (Freitag et al., 2021a), with a bigger emphasis set to sentence-level prediction.[9]

Interpretation of the Curriculum. We start by training a sentence-level metric—similar to UNITE (Wan et al., 2022a)—on the vastly available DA annotations. Phase I acts as a warm-up for subsequent stages. In fact, prior research has shown that models trained on DA annotations leverage token-level information that aligns with MQM error annotations (Rei et al., 2023b). Moving to Phase II, we assume we have a metric that can perform sentence-level regression. Thus, the aim here shifts to integrating word-level supervision without compromising the previously acquired sentence-level prediction skills. To do so, we use the highly diverse corpora of MQM annotations and set most emphasis on the word-level task. Finally, we exclusively leverage a small corpus (around 25k samples) of very high-quality MQM annotations from (Freitag et al., 2021a)—each sample has three annotations

[6] We measure cross-lingual similarity using sentence embeddings obtained with the LaBSE encoder (Feng et al., 2022).

[7] This is particularly relevant for DA annotations, since these judgments typically come from non-expert annotators.

[8] This was initially introduced in BLEURT-20 (Pu et al., 2021).

[9] The achieved λ weights for Phases II and III were $\lambda = 0.983$ and $\lambda = 0.055$, respectively.

from separate annotators—with additional synthetic hallucinations. Our focus here is to mitigate any potential decline in sentence-level regression capabilities during Phase II.

5 Experimental Setting

5.1 Evaluation

We evaluate the performance of our metrics using two datasets: (i) the MQM annotations from the News domain of the WMT 2022 Metrics shared task, and (ii) the WMT 2023 Metrics shared task evaluation suite. The WMT22 annotations encompass three language pairs: Chinese→English (`zh-en`), English→German (`en-de`), and English→Russian (`en-ru`). On the other hand, the WMT23 annotations cover Chinese→English (`zh-en`), English→German (`en-de`), and Hebrew→English (`he-en`). We evaluate the metrics in terms of sentence-level, system-level, and error span prediction performance.

At the sentence-level, we report Kendall's Tau (τ) using the Perm-Both hypothesis test (Deutsch et al., 2021). We also evaluate the metrics on System-level Pairwise Accuracy (Kocmi et al., 2021). We base these evaluations on 200 re-sampling runs, with a significance level (p) set to 0.05. For error span prediction, we adopt the WMT23 Quality Estimation shared task evaluation methodology and compute F1 scores calculated at the character level, taking into account partial matches for both minor and major errors.[10] For WMT23, we follow the evaluation setup from the shared task (Freitag et al., 2023) and report the aggregated System-level Pairwise Accuracy pooled across all language pairs, and the primary metric Average Correlation, which encompasses ten tasks, spanning system- and sentence-level metrics.[11]

5.2 Baselines

Sentence and System-level. We test our metrics against widely used *open* neural metrics: COMET-22 (Rei et al., 2022a) and BLEURT-20 (Pu et al., 2021). Additionally, we include METRICX, the best performing metric from the WMT22 Metrics shared task (Freitag et al., 2022),[12] and GEMBA (Kocmi and Federmann, 2023b), which employs GPT4 (OpenAI, 2023) to evaluate translations following DA guidelines. For WMT23, we report the same metrics but update them, when needed (for METRICX-23 [Juraska et al., 2023] and GEMBA-MQM [Kocmi and Federmann, 2023a]), with the versions submitted to the official competition.[13]

Error Span Prediction. We report results using GPT3.5 and GPT4 models, by prompting it in the style of AUTOMQM (Fernandes et al., 2023).[14] We carefully select 5 shots that are held constant for all samples. This way, we can directly compare our results with state-of-the-art LLMs, which have been shown to be able to perform the task of error detection (Fernandes et al., 2023; Xu et al., 2023).

6 Correlations with Human Judgments

In this section, we present a standard performance analysis of our metrics in terms of correlations with human judgments. Overall, we find XCOMET to be a state-of-the-art in sentence-level and error span prediction, being competitive with generative LLMs in terms of system-level evaluation.

Sentence-level Evaluation. Table 2a shows that both XCOMET metrics outperform other strong performing neural metrics, including the generative approach leveraging GPT4 of GEMBA. In particular, XCOMET-XXL sets a new state-of-the-art for `en-de` and `en-ru`. Interestingly, we can see that, while scaling up the encoder model of the XCOMET metrics (from XL to XXL) holds better results, XCOMET-XL is very competitive. In fact, it outperforms METRICX, which runs at even a larger size than XCOMET-XXL. Finally, we can also observe that the MQM scores inferred exclusively from the predicted error spans also exhibit strong performance, outperforming widely used metrics BLEURT-20 and COMET-22. This is particularly relevant: the predicted error spans bring not only a

[10] We convert all critical errors into major errors, in order to match the guidelines described in Freitag et al. (2021a) that were used for annotating the `zh-en` and `de-en` test sets.

[11] The ten tasks consist of the System-level Pairwise Accuracy pooled across the language pairs, as well as segment-level pairwise ranking accuracy with tie calibration (Deutsch et al., 2023a), and system- and segment-level Pearson correlation for each of the individual language pairs.

[12] Specifically, we employ the `metricx_xxl_MQM_2020` submission scores from the `mt-metrics-eval` package. Although the metric has not been released publicly, it is public that it is built upon the mT5-XXL (Xue et al., 2021) and has 13B parameters (Deutsch et al., 2023b).

[13] For all baselines, we report the official numbers from the WMT23 Metrics Shared Task (Freitag et al., 2023).

[14] We use the models from the OpenAI API (`gpt-3.5-turbo` and `gpt-4`) in October 2023.

Metric	zh-en	en-de	en-ru	Avg.
BLEURT-20	0.336	0.380	0.379	0.365
COMET-22	0.335	0.369	0.391	0.361
METRICX	**0.415**	0.405	0.444	0.421
GEMBA-GPT4-DA	0.292	0.387	0.354	0.354
XCOMET-XL	0.399	0.414	0.448	0.421
XCOMET-XXL	0.390	**0.435**	**0.470**	**0.432**
MQM scores from the error spans ($\hat{y} = \hat{y}_{\text{MQM}}$)				
XCOMET-XL (MQM)	0.374	0.389	0.445	0.402
XCOMET-XXL (MQM)	0.332	0.415	0.439	0.395

(a) Sentence-level evaluation.

Metric	zh-en	en-de	en-ru	Avg.
BLEURT-20	0.762	0.771	0.743	0.759
COMET-22	0.705	0.800	0.733	0.746
METRICX	0.762	0.781	0.724	0.756
GEMBA-GPT4-DA	0.752	**0.848**	**0.876**	**0.825**
XCOMET-XL	**0.800**	0.743	0.790	0.778
XCOMET-XXL	**0.800**	**0.829**	**0.829**	**0.819**
MQM scores from the error spans ($\hat{y} = \hat{y}_{\text{MQM}}$)				
XCOMET-XL (MQM)	0.781	0.762	0.762	0.768
XCOMET-XXL (MQM)	0.781	**0.838**	0.810	0.810

(b) System-level evaluation.

Table 2: Segment-level Kendall-Tau (↑) in (a), and System-level Pairwise Accuracy (↑) in (b) using the Perm-Both hypothesis test (Deutsch et al., 2021) on the WMT22 Shared Task News domain test set. Numbers in bold belong to the top-performing cluster according to statistical significance ($p < 0.05$).

more detailed view into translation errors but also provide high-quality sentence-level scores.

System-level Evaluation. Table 2b and Table 3 show results for system-level for both WMT22 and WMT23 test sets. Similarly to what we observed at the sentence-level, our metrics show consistently superior performance when compared to other dedicated neural metrics. Notably, although generative approaches typically do much better at system-level evaluation when compared to dedicated models (Kocmi and Federmann, 2023b; Fernandes et al., 2023), XCOMET-XXL remains competitive in all language pairs with GEMBA using GPT4. Finally, building on the findings at the sentence-level, Table 2b reveals that the MQM scores inferred directly and exclusively from the predicted error spans also exhibit very competitive performance in terms of system-level accuracy.

Aggregated Evaluation. Table 3 shows aggregated results for the WMT23 Metrics Shared Task. Our metrics, at both scales, would win the shared task, outperforming both the newest version of METRICX and GEMBA-MQM. Following the trend presented at sentence-level evaluation, we note that XCOMET-XL is indeed competitive with XCOMET-XXL although running at a smaller scale.

Error Span Prediction. While we have highlighted the utility of the predicted error spans through the inferred sentence-level MQM scores, here we turn to evaluating them directly. Table 4 shows that the error spans predicted via XCOMET metrics outperform those obtained with both GPT3.5 and GPT4 despite being smaller in capacity relative to these models. In fact, our metrics achieve close performance to that of GPT4, even when a reference is not provided.

Metric	system-level acc.	avg-corr.
BLEURT-20	0.892	0.776
COMET-22	0.900	0.779
METRICX-23	0.908	0.808
GEMBA-MQM	**0.944**	0.802
XCOMET-XL	0.912	**0.813**
XCOMET-XXL	0.920	**0.812**

Table 3: System-level pairwise accuracy (↑) (Kocmi et al., 2021) computed over data pooled across all three WMT23 language pairs, and primary metric Average Correlation (↑).

Metric	zh-en	en-de	en-ru	Avg.
● AutoMQM (GPT3.5)	0.143	0.160	0.166	0.156
● AutoMQM (GPT4)	0.248	0.257	**0.281**	0.262
● XCOMET-XL	0.237	0.290	**0.281**	0.269
● XCOMET-XXL	**0.257**	**0.320**	0.262	**0.280**
Error spans detected with source-only input				
● XCOMET-XL (SRC)	0.208	0.264	0.252	0.242
● XCOMET-XXL (SRC)	0.229	0.298	0.238	0.255

Table 4: F1 scores (↑) for error span detection: reference-free (●), reference-based (●) evaluation.

Interplay of Error Spans and Sentence-level Scores. Table 5 shows a strong correlation between the different score types predicted by XCOMET and the MQM inferred score derived exclusively from error spans. This interplay is highly important: the predicted error spans may be valuable, not just for the sake of accuracy but also for interpretability. Interestingly, these high correlations with the predicted scores from each forward pass ($\hat{y}_{\text{SRC}}$, $\hat{y}_{\text{REF}}$, $\hat{y}_{\text{SRC+REF}}$) are obtained despite no explicit alignment mechanism governing the relationship between the predictions of the

SCORE	zh-en	en-de	en-ru	All
$\hat{y}_{\text{SRC}}$	0.73	0.75	0.79	0.78
$\hat{y}_{\text{REF}}$	0.75	0.74	0.75	0.77
$\hat{y}_{\text{SRC+REF}}$	0.78	0.79	0.82	0.82
$\hat{y}^{\dagger}_{\text{SL}}$	0.90	0.92	0.92	0.92

Table 5: Pearson correlations between the regression scores produced by **XCOMET-XXL** ($\hat{y}_{\text{SRC}}$, $\hat{y}_{\text{REF}}$, $\hat{y}_{\text{SRC+REF}}$, $\hat{y}_{\text{SL}}$) and the MQM inferred score, $\hat{y}_{\text{MQM}}$, computed from the identified error spans. †The computation of $\hat{y}_{\text{SL}}$, contrary to the other scores, makes direct use of $\hat{y}_{\text{MQM}}$ (see Eq. 7).

sentence-level and word-level heads. We hypothesize that it is thus the shared encoder that, during the multi-task training, aligns the representations between the two tasks. As such, XCOMET provides, through its predicted error spans, a potential lens through which we can better understand, contextualize, and even debug its own sentence-level predictions.

7 Robustness of XCOMET to Pathological Translations

We have shown that XCOMET metrics exhibit state-of-the-art correlations with human judgements when evaluating on high-quality MQM annotations. However, these MQM annotations are often highly unbalanced and contain little to no major or critical errors. As such, they may not offer a full picture of the metrics' performance. In this section, we shift our focus to studying how XCOMET metrics behave when evaluating translations with localized major or critical errors, and highly pathological translations, such as hallucinations.

7.1 Localized Errors

We employ SMAUG (Alves et al., 2022),[15] a tool designed to generate synthetic data for stress-testing metrics, to create corrupted translations that contain major or critical errors. We generate translations with the following pathologies: addition of text, negation errors, mask in-filling, named entity errors, and errors in numbers. For this evaluation, we use data from the WMT 2023 Metrics shared task. Specifically, we corrupt the released *synthetic* references for which the XCOMET metrics found no errors.[16] Moreover, as the full suite of SMAUG transformations can only be applied to English text, we focus on Chinese→English (zh-en) and Hebrew→English (he-en) translations.

	zh-en		he-en	
ERROR	XL	XXL	XL	XXL
Add. of text	3.66	10.7	6.15	7.35
Negation	0.20	0.20	3.89	4.90
Mask in-fill	5.01	17.0	4.78	3.92
Swap NUM	3.19	2.88	0.16	0.00
Swap NE	3.66	6.94	9.81	7.01
All	**2.24**	10.7	9.81	**7.00**

Table 6: Percentage (%) of translations, segmented by perturbation type, that are predicted to have no errors (↓). We show results for both zh-en and he-en language pairs across XCOMET sizes.

XCOMET Predicts most Localized Errors as Major or Critical Errors. Table 6 shows that XCOMET metrics identify errors in the vast majority of the perturbed samples, with trends varying across scale and language pair. We found that the errors predicted by XCOMET-XL and XCOMET-XXL overlap with the artificially induced perturbations in over 90% of the perturbed samples (98% for XL and 90.9% for XXL). However, upon further analysis of XCOMET's predicted error spans, we observed that the model tends to identify additional spans in the perturbed sentence as erroneous, beyond the induced perturbations. This behavior is more prominent for perturbations involving the addition of text, which are not as localized as perturbations like swapping numbers or named entities. Furthermore, we noticed that the model has a propensity to assign the same error category to all predicted spans within a single sentence. When the metric predicts multiple error spans in a sentence, it assigns different severity levels to those spans only about 35% of the time. Improving the model's ability to differentiate error categories among multiple errors within a sentence is an interesting avenue for future research

[15] https://github.com/Unbabel/smaug.

[16] This allows us to isolate the effect of the perturbations. In case there are predicted error spans for the transformed translations, these are a result of the perturbation induced.

Source: 最后，我们设定了两个大的战略方向。
Translation: In the end, we set two major strategic directions.
Perturbed Translation: (`Negation error`) In the end, we `PERT:` don't even see two major strategic directions.
xCOMET error span predictions: In the end, `CRIT:` we don't even see two major strategic directions.
Source: 海地国内的一些地区，多达90%的房屋在地震中被摧毁。
Translation: In some areas of Haiti, as many as 90% of houses were destroyed in the earthquake.
Perturbed Translation: (`Swap NUM`) In some areas of Haiti, as many as `PERT:` 37.5% of houses were destroyed in the earthquake.
xCOMET error span predictions: In some areas of Haiti, as many as `MAJ:` 37.5% of houses were destroyed in the earthquake.
Source: 1993 年，莫德罗在 1989 年 5 月的市政选举中被判选举舞弊罪名成立，但并未入狱。
Translation: In 1993, Modrow was convicted of electoral fraud in the May 1989 municipal elections, but did not go to prison.
Perturbed Translation: (`Mask in-fill`) In 1993, Modrow was convicted of electoral fraud in the `PERT:` May 1989 presidential election, but did not go to prison.
xCOMET error span predictions: In 1993, Modrow was convicted of electoral fraud in the May 1989 `MAJ:` presidential election, but did not go to prison.

Table 7: Predictions of XCOMET-XL for perturbed translations. We highlight minor (`MIN`), major (`MAJ`), and critical (`CRIT`) error spans.

and development. We show several examples of predictions of XCOMET in Table 7.

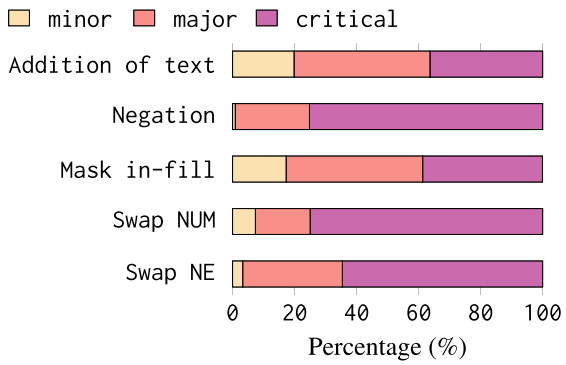
We also found that negation errors and mismatches in numbers are the most easily identified by the metrics. This is interesting: Localized errors, such as mismatches in numbers and named-entity errors, had been pinpointed as weaknesses of previous COMET metrics (Amrhein and Sennrich, 2022; Raunak et al., 2022). This earlier limitation seems to now have been addressed successfully. In fact, the results in Figure 4a show that most of these errors are predicted as critical errors. One plausible hypothesis for these improvements is the incorporation of datasets that contain negative translations and synthetic hallucinations into our training set.

XCOMET Sentence-level Scores Are Sensitive to Localized Perturbations. Figure 4b shows that localized errors can lead to significant decreases in the predicted sentence-level scores, with perturbation-wise trends mirroring those of the error span predictions: the most pronounced decreases are found for negation errors and mismatches in numbers and named-entities (median decreases of around 20 points). The distribution of the decreases in quality also reveals two relevant trends: (i) localized perturbations can cause XCOMET-XXL to shift from a score of a perfect translation to that of an unrelated translation, and (ii) the behavior of XCOMET-XXL is not perfect and can be further improved: In rare cases, perturbations may actually lead to an increase in the score. Nevertheless, upon closer inspection, for over 90% of cases, the increase is smaller than 1 point.

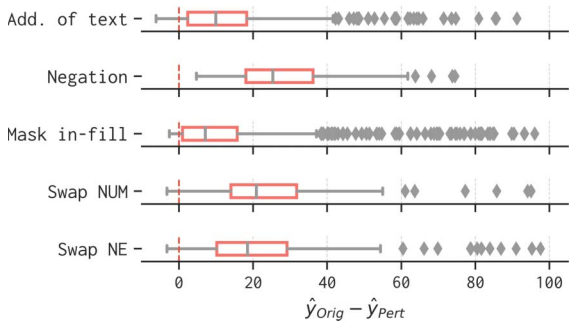
7.2 Hallucinations

Hallucinations lie at the extreme-end of machine translation pathologies (Raunak et al., 2021), and can have devastating impact when models are deployed *in the wild*. Yet, these translations are often overlooked when assessing the performance of translation systems. Their rarity means that performance, usually judged according to an aggregated corpus-level score, may remain largely unperturbed by a very small number of hallucinations. Here, we assess how the XCOMET metrics rank hallucinations among other translations. We will use the German→English hallucination benchmark introduced in Guerreiro et al. (2023). This benchmark involves over 3.4k translations—produced by an actual machine translation system—of different error types, including omissions, named-entity errors, and hallucinations (oscillatory, fully, and strongly detached). For a metric that has not been trained explicitly to rank translations, the benchmark is quite challenging: Hallucinations should be ranked below other severe errors and incorrect translations.

XCOMET Metrics Can Distinguish Hallucinations from Other Translations. The results in Table 8 show that both XCOMET metrics largely rank hallucinations lower than other errors. This is especially true for the most severe type of hallucination (fully detached), for which the AUROC exceeds 95 for the XXL metric. In fact, Figure 5 reveals that XCOMET-XXL assigns over 90% of these fully detached hallucinations a score under 10. We show examples of error spans predicted



(a) Percentage of error types on data with critical errors (for both zh-en and he-en data), as predicted by **XCOMET-XXL**.



(b) Impact of the perturbations, as measured by the difference in **XCOMET-XXL** ($\hat{y} = \hat{y}_{SL}$) between the original and the perturbed translation, on the zh-en data.

Figure 4: Analysis of XCOMET-XXL for data with localized critical errors in terms of (a) distribution of error severities for the predicted error spans, and (b) sensitivity of the sentence-level scores.

METRIC	All	Full Det.	Osc.
• BLEURT-20	0.824	0.892	0.799
• COMET-22	0.829	0.878	0.883
• COMETKIWI-XXL	0.839	0.834	0.902
• **XCOMET-XL**	0.865	0.907	0.922
• **XCOMET-XXL**	0.890	**0.964**	0.844
QE scores from the error spans ($\hat{y} = \hat{y}_{SRC}$)			
• **XCOMET-XL** (SRC)	0.885	0.924	**0.944**
• **XCOMET-XXL** (SRC)	**0.902**	0.959	0.866

Table 8: Hallucination detection performance on the de-en hallucination benchmark from Guerreiro et al. (2023) as measured by AUROC (↑) for reference-free (•) and reference-based (•) metrics. We report results for all the dataset, for fully detached, and oscillatory hallucinations separately.

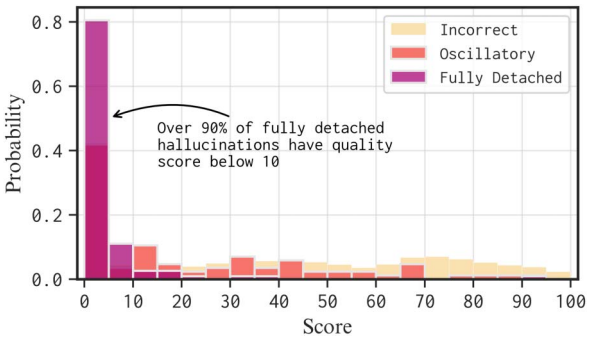


Figure 5: Category-wise distribution of **XCOMET-XXL** scores on the hallucination benchmark.

Source:
Das Teilabonnement für international tätige Juristen.

Translation:
The sub-sub-sub-sub-scription for international lawyers.

Reference:
Partial subscriptions for internationally active lawyers.

Oscillatory hall. with XCOMET error span predictions:
The **CRIT:** sub-sub-sub-sub-scription for international lawyers.

Source:
Empfehlenswert gleich mit der Zimmerreservierung zu buchen!

Translation:
The staff were very friendly and helpful. The room was clean and comfortable.

Reference:
We recommend booking your treatments together with the hotel booking!

Fully detached hall. with XCOMET error span predictions:
CRIT: The staff were very friendly and helpful. The room was clean and comfortable.

Table 9: Examples of predictions of XCOMET-XXL for the hallucination data of Guerreiro et al. (2023). The model correctly identifies the anomalous repeated phrase for the oscillatory hallucination, and predicts the whole translation as a single error span in the case of the fully detached hallucination.

by XCOMET-XXL in Table 9. Relative to previous metrics, XCOMET achieves overall improvements. Interestingly, we also find that SRC-based evaluation (i.e., without the use of a reference) can reap benefits in this scenario. We hypothesize that this is due to the metric over-relying on the reference when it is available (Rei et al., 2023b). While hallucinations contain content that is detached from the source, some of their text may still overlap (even if just lexically) the reference (e.g., in strongly detached or oscillatory hallucinations), leading to higher scores.

8 Ablations on Design Choices

We now address relevant questions about the development of XCOMET through ablations on design choices. Ablations are run with XCOMET-XL on the

	zh-en		en-de		en-ru		Avg.	
STAGE	τ	F1	τ	F1	τ	F1	τ	F1
*Post Phase I with **sentence-level only** objective: $\lambda = 0$*								
PHASE **I**	0.377	NA	0.356	NA	0.425	NA	0.386	NA
PHASE **II** ($\hat{y} = \hat{y}_{\text{REG}}$)	0.372	NA	0.386	NA	0.448	NA	0.402	NA
PHASE **III** ($\hat{y} = \hat{y}_{\text{REG}}$)	0.395	NA	0.391	NA	0.457	NA	0.414	NA
*Post Phase I with **word-level only** objective: $\lambda = 1$*								
PHASE **II** ($\hat{y} = \hat{y}_{\text{REG}}$)	0.333	0.293	0.357	0.332	0.415	0.229	0.368	0.285
PHASE **II** ($\hat{y} = \hat{y}_{\text{MQM}}$)	0.331	=	0.410	=	0.395	=	0.379	=
*Post Phase I with **multi-task only** objective: λ set as described in Section 4.4*								
PHASE **II** ($\hat{y} = \hat{y}_{\text{MQM}}$)	0.330	0.284	0.359	0.328	0.413	0.212	0.367	0.275
PHASE **II** ($\hat{y} = \hat{y}_{\text{SL}}$)	0.368	=	0.396	=	0.420	=	0.395	=
PHASE **III** ($\hat{y} = \hat{y}_{\text{MQM}}$; XCOMET)	0.374	0.237	0.389	0.290	0.445	0.281	0.402	0.269
PHASE **III** ($\hat{y} = \hat{y}_{\text{SL}}$; XCOMET)	0.399	=	0.597	=	0.448	=	0.421	=

Table 10: Segment-level Kendall-Tau (τ) (↑) and F1 scores (↑) for error span detection on different curriculum choices. We represent metrics that can *only* perform segment-level evaluation, *only* perform word-level evaluation ($\lambda = 1$), and both. When a model has no capabilities to perform error span prediction, we write NA under its F1 score.

MQM annotations from the News domain of the WMT 22 Metrics shared task (see Section 5).

Impact of the Training Curriculum. We employed a curriculum to train XCOMET (see Section 4.4) in order to balance data from different annotation strategies (i.e., DA and MQM annotations), and also to better balance the multi-task objective. Here, we want to assess how performance evolves throughout the different stages. We perform ablations on Phase II and Phase III, which correspond to the introduction of the multi-task objective and MQM training data that contain both sentence-level scores and error spans.

Table 10 shows that while a multi-task model outperforms single-task models for sentence-level evaluation,[17] it does not hold true for word-level evaluation. The best word-level model is obtained by doing Phase II with a word-level only objective. Note that we can still extract sentence-level scores from such a model in two ways: (i) by leveraging the still existing regression head trained during Phase I, or (ii) by converting the error spans into a single sentence-level score. However, notably, neither of these approaches is competitive with our final XCOMET model. In fact, it turns out, performing sentence-level evaluation via the error spans predicted by the final model leads to better correlations than with the word-level only model. Moreover, aggregating the different scores from the regression heads yields the best overall performance for sentence-level evaluation.

[17]For sentence-level only models, we present the sentence-level score $\hat{y} = \hat{y}_{\text{REG}}$ correspondent to setting uniform weights across all three individual scores (SRC, REF, and SRC+REF).

SCORE	zh-en	en-de	en-ru	All
Individual Scores				
$\hat{y}_{\text{SRC}}$	0.368	0.358	0.402	0.376
$\hat{y}_{\text{REF}}$	0.399	0.389	0.427	0.405
$\hat{y}_{\text{SRC+REF}}$	0.399	0.390	0.438	**0.409**
$\hat{y}_{\text{MQM}}$	0.374	0.389	0.445	0.402
Aggregated Regression Scores: see Equation 7				
$\hat{y}_{\text{REG}}$	0.402	0.380	0.4401	0.408
$\hat{y}_{\text{UNIF}}$	0.398	0.402	0.448	0.416
$\hat{y}_{\text{SL}}$	0.399	0.414	0.448	**0.421**

Table 11: Segment-level Kendall-Tau (τ) (↑) for individual and aggregated sentence-level scores.

Impact of the Weights on the Sentence-level Scores from Equation (7). We studied the impact of aggregating different sentence-level scores by varying the weights $\boldsymbol{w}$ in Equation 7. Besides the individual scores for each evaluation mode (SRC, REF, and SRC+REF), we present three aggregations: (i) $\hat{y} = \hat{y}_{\text{SL}}$ used in the final model, (ii) $\hat{y} = \hat{y}_{\text{REG}}$ with uniform weights across the three individual scores and not considering the inferred MQM score from the error spans ($\boldsymbol{w} = [1/3, 1/3, 1/3, 0]$), and (iii) $\hat{y} = \hat{y}_{\text{UNIF}}$ with uniform weights across all scores ($\boldsymbol{w} = [1/4, 1/4, 1/4, 1/4]$).

Table 11 reveals two interesting findings: (i) aggregating scores does not always outperform

individual scores (e.g., $\hat{y}_{\text{REG}}$ performs similarly to $\hat{y}_{\text{SRC+REF}}$), and (ii) including an inferred MQM score obtained through error span prediction boosts sentence-level performance. Notably, the improvement of our final aggregated score over $\hat{y}_{\text{SRC+REF}}$ is not substantial. This suggests that, under computational constraints, one could consider computing a single $\hat{y}_{\text{SRC+REF}}$ score without the need for three different forward passes and error span prediction.

9 Conclusions

We introduced XCOMET, a novel suite of metrics for machine translation evaluation that combines sentence-level prediction with fine-grained error span prediction. Through extensive experiments, we have shown that XCOMET is a state-of-the-art metric at all relevant vectors of evaluation: sentence-level, system-level, and error span prediction. Notably, through XCOMET's capabilities to predict error spans, we can not only obtain useful signals for downstream prediction (either directly through error span prediction or by informing sentence-level scores) but also gain access to a lens through which we can better understand and interpret its predictions. We also stress-tested the metrics by assessing how they score localized critical errors and hallucinations: The metrics identify the vast majority of localized errors and can appropriately penalize the severity of hallucinations.

We hope XCOMET can serve as a step towards more informed machine translation evaluation.

Acknowledgments

We are grateful to José Pombal, José G. C. de Souza, and Sweta Agrawal for their valuable feedback and discussions.

This work was supported by the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI, by the European Research Council (DECOLLAGE, ERC-2022-CoG 101088763), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), and by the Fundação para a Ciência e Tecnologia (contracts UIDB/50021/2020 and UIDB/50008 /2020). We also thank the HPC resources from GENCI-IDRIS (grants 2023–AD011014714, 2023–AD0110146A68R1, and AD011012377R2).

References

Duarte Alves, 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, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Keqin Bao, Yu Wan, Dayiheng Liu, Baosong Yang, Wenqiang Lei, Xiangnan He, Derek F. Wong, and Jun Xie. 2023. Towards fine-grained information: Identifying the type and location of translation errors.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4717

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146. https://doi.org/10.1162/tacl_a_00417

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

Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.798

Daniel Deutsch, Juraj Juraska, Mara Finkelstein, and Markus Freitag. 2023b. Training and meta-evaluating machine translation evaluation metrics at the paragraph level. In *Proceedings of the Eighth Conference on Machine Translation*, pages 994–1011, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.96

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.62

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. 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 1064–1081, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.100

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.100

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5401

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474. https://doi.org/10.1162/tacl_a_00437

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.51

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

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

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.repl4nlp-1.4

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.75

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. 2023. Reinforced self-training (rest) for language modeling.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.63

Marzena Karpinska, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.649

Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.64

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

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

Christoph Leiter, Piyawat Lertvittayakumjorn, M. Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. 2023. Towards explainable evaluation metrics for machine translation. *ArXiv*, abs/2306.13041.

Arle Lommel, 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. https://doi.org/10.5565/rev/tradumatica.77

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.566

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073135

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

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-3049

Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for MT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 751–762, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.58

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.92

Vikas Raunak, Matt Post, and Arul Menezes. 2022. SALTED: A framework for SAlient long-tail translation error detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5163–5179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.379

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023a. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 839–846, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.73

Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. 2023b. 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, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-short.94

Ricardo Rei, 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. Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of*

the Seventh Conference on Machine Translation (WMT), pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ananya Sai B., Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. 2023. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.795

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.704

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24:39–50. https://doi.org/10.1007/s10590-010-9077-2

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Dazhen Wan, Zheng Zhang, Qi Zhu, Lizi Liao, and Minlie Huang. 2022a. A unified dialogue user simulator for few-shot data augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3788–3799, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.277

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022b. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.558

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. Instructscore: Towards explainable text generation evaluation with automatic feedback.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.41

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.