Meta-Evaluation of Automatic Machine Translation Metrics between Italian and a Minor Language Variety of German

Paolo Di Natale[1,*], Elena Chiocchetti[1] and Egon Stemle[1,2]

[1]Eurac Research, Viale Druso/Drususallee 1, 39100 Bolzano/Bozen, Italy
[2]Masaryk University, Zerotinovo namesti 9, 602 00 Brno, Czech Republic

Abstract

We present the first meta-evaluation of Automatic Machine Translation Evaluation (AMTE) metrics between Italian and South Tyrolean German, a low-resourced standard variety of German. This minor German variety is recognised as a co-official language at the local level and is used by the local public administration and legislature. We evaluate metric agreement with human judgement across translation quality levels, using a dataset of bilingual machine-translated decrees annotated with human-curated error tags. Our findings show that embedding-based metrics perform best for evaluating high-quality translations, while learned neural metrics correlate more strongly with human judgments on lower-quality ranges. We also expose a persistent bias in AMTE against minor language varieties and make suggestions about the design of linguistic resources for envisaged custom metric devolopment.

Keywords

automatic machine translation evaluation metrics, metrics meta-evaluation, non-English language combination, minor language variety, machine translation, natural language generation evaluation, specialized communication

1. Introduction

South Tyrolean German is a minor standard variety of German with a co-official status in the Italian province of Bolzano/Bozen (South Tyrol). The 350,000 German-speaking citizens in South Tyrol have the right to communicate with and access public services in their native language at the local level. Given the increasing integration of AI technologies into everyday life, this context underscores the need of developing bilingual NLP tools tailored to the South Tyrolean variety of German and use cases, with Machine Translation (MT) one of the most pressing fields of research. However, it is well documented that the performance of NLP systems for minor language varieties significantly lags behind both their major counterparts and high-resource languages [1].

Interest in generating translations into minor language varieties is growing, yet the lack of validated evaluation metrics hampers accurate monitoring of achieved progress. Most related studies still rely on inadequate, superseded lexical-overlap methods [2]. While the research community has made efforts to adapt neural metrics for under-resourced and dialectal varieties [3, 4], the development of robust evaluation methods is complicated by the absence of high-quality, sufficiently large labeled datasets – an issue common to all under-resourced varieties [5]. Knowles et al. [6] have called for a comparative evaluation, as they argue that metrics assign lower scores to minor lexical variants even when no change in meaning exists. In addition, inefficient tokenization methods lead to suboptimal segmentation and reduced adaptability for under-resourced languages [7].

Prior experiments with adaptive MT for South Tyrol [8, 9] have also employed metrics based on lexical overlap despite their known underperformance compared to neural metrics. This reliance stems from the lack of a thorough, localized evaluation of more advanced metric paradigms and makes a compelling case for a dedicated meta-evaluation study of existing solutions applicable to the South Tyrolean context.

This work presents the first such MT meta-evaluation study of metrics for the Italian–South Tyrolean German language pair. We conduct our analysis on MT@BZ[1], a manually error-annotated corpus of legal texts covering both translation directions, to assess the reliability of current automatic evaluation metrics.

1.1. Automatic Machine Translation Evaluation

Human evaluation remains the gold standard method for assessing MT quality outputs. However, because human annotation is time-consuming, resource-intensive and requires high domain expertise, Automatic Machine Translation Evaluation (AMTE) metrics have garnered increasing attention. These metrics aim to estimate translation quality by comparing a system-generated *candi-*

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
paolo.dinatale@eurac.edu (P. D. Natale);
elena.chiocchetti@eurac.edu (E. Chiocchetti);
egon.stemle@eurac.edu (E. Stemle)

[1]http://hdl.handle.net/20.500.12124/60

date translation either to the *source* segment[2] in the other language, to a human-produced *reference* translation in the same language, or to both. In scenarios where the output of only one translation system is available, as in this case, a so-called segment-level evaluation is carried out. It consists of "evaluating metrics based on their ability to rank segments in the same order as human judgments" [10]. The effectiveness of such metrics is commonly measured using ranking correlation coefficients, under the assumption that a reliable metric should consistently assign higher scores to translations deemed superior by human annotators [11].

Existing metrics can be categorized into three main types:

- **String-based**: this approach quantifies translation quality by measuring lexical overlap with one or more reference translations. These methods operate at the surface level, comparing exact matches of word or character sequences between the candidate and the reference.
- **Embedding-based**: these metrics leverage contextualized token embeddings from pretrained language models to compute semantic similarity between the candidate translation and the reference. Semantic alignment is evaluated at the token level using cosine similarity, followed by an F-score aggregation procedure.
- **Learned**: these metrics are based on transformer architectures that have been fine-tuned via supervised learning to replicate human judgments of machine translation quality, typically using a regression objective to provide a continuous score.

2. Motivation

2.1. Social and Linguistic Background of South Tyrol

South Tyrolean German is the standard variety of German used in the Autonomous Province of Bolzano/Bozen (South Tyrol) in Northern Italy. In South Tyrol, German is a recognized minority language, co-official with Italian. Public administration offices are legally required to use German when interacting with the German-speaking population (Presidential Decree No. 670/1972, Art. 100), which makes up the large majority of South Tyrol's population (69%)[3]. Consequently, all administrative documents, local legislation, and materials intended for the general public – such as the websites of local public institutions – must be available not only in the national language Italian but also in the minority language German[4].

This multilingual institutional language regime is largely implemented through translation between Italian and German or vice-versa. National legislation is drafted in Italian and any implementations at local level create the need for translation into German. Following quotas in public employment, about two thirds of public administration staff is German-speaking. Consequently, many legal and administrative texts are now originally drafted in German. While the Italian and German version of, for example, a local law are both official, in case of diverging interpretation the Italian version prevails (Presidential Decree No. 670/1972, Art. 99). This means that a translated text can become the legally binding version. Given the growing use of machine translation, this holds true also for machine-translated or post-edited texts.

The impressive level of fluency of MT-generated texts poses a challenge for fair quality assessment of MT systems even for human evaluators – especially for those lacking specialized training, who may be outperformed by automated neural metrics [12]. In South Tyrolean public offices, where translation-related tasks are often performed by non-specialists, the rising adoption of MT – frequently without adherence to scientific evaluation protocols [13] – carries the risk of overestimating productivity gains. Without systematic, targeted performance monitoring, critical errors may go unnoticed. As highlighted in the error analysis of a machine-translated legal corpus [14], MT systems often struggle with local legal terminology and are prone to interference from other legal systems using German. For example, *kommunale Steuer* (municipal tax) is never used in South Tyrol as it would in Germany. The South Tyrolean term for "municipal tax" is *Gemeindesteuer*. Such errors can severely compromise translation quality and usability. In high-stakes domains like the legal one, fluency is secondary to semantic precision and legal appropriateness. Critical accuracy errors can distort meaning, making translated laws unpublishable or even harmful. Consequently, there is a clear need for MT evaluation frameworks that attend to the specific requirements of the South Tyrolean administration and population.

2.2. Toward the Development of Custom Metrics

The well-documented challenges of adapting NLP applications to minor language varieties [1] also apply to the

[2] In the field of MT, a *segment* is defined as the minimal translation unit, which in this study corresponds to a sentence.

[3] See the latest census data: https://assets-eu-01.kc-usercontent.com/b5376750-8076-01cf-17d2-d343e29778a7/5deec178-b2a3-4e2d-8795-d37635c7e0f7/pressnote_1160209_mit56_2024.pdf

[4] There is a third official language, Ladin, spoken by about 20,000 South Tyroleans. We will not deal with Ladin in this paper.

development of automatic evaluation metrics. Language models are pre-trained onto large-scale corpora where major language varieties contribute a disproportionately larger amount of training signal [15], often without explicit annotation of variety or dialect tag. This results in biased representations and undermines the fairness and reliability of evaluation metrics for underrepresented varieties [16]. Current literature has shown that intensive continued pre-training [16] and the use of high-quality, human-annotated datasets spanning a range of translation quality levels [4] are essential to improving evaluation performances. Yet, these strategies remain largely impractical at present due to the significant data and resource demands they entail.

Also, given the high costs of constructing fine-grained, manually annotated datasets, one wants to be sure that the compilation of structured and detailed linguistic resources is empirically justified. While Amrhein et al. [17] argue that the inclusion of reference translations generally improves evaluation reliability, the behavior of existing metrics remains inconsistent, occasionally even counterintuitive. For example, some metrics have been observed to disregard the reference altogether [18], or to produce high scores even when the source text is omitted entirely [6, 10]. As a result, a comprehensive assessment of existing solutions is needed not only in terms of the identification of the best suited metrics to the context under study, but also to lay the groundwork for envisaged future metric development.

Moreover, reliable metrics can also advance generation tasks. An emerging trend of natural language generation is to exploit Minimum Bayes Risk (MBR) decoding, which selects the output hypothesis that minimizes expected loss according to a utility function defined by a chosen evaluation metric [19]. This approach can act as a form of style transfer with a reduction in training costs and data requirements. However, using the same metric for both decoding and final evaluation introduces bias, as the system is optimized to reproduce the metric's idiosyncrasies [20]. Even different but highly correlated metrics – especially if they are of the same type – can produce similar biases [21]. Thus, evaluating the robustness of multiple metric paradigms becomes an essential prerequisite to generating text in South Tyrolean German with MBR decoding.

3. Challenges of Automatic Machine Translation Evaluation

Learned metrics have consistently outperformed other evaluation methods in benchmark competitions such as the WMT Metrics Shared Task [22]. However, this finding should not be generalized uncritically. Since neural metrics are predominantly fine-tuned on WMT competition datasets – which represent a limited range of linguistic diversity and domains – their superiority in more specialized evaluation scenarios remains open to question.

Knowles et al. [6] raise questions regarding how metrics assess terminological variation within language varieties and call for more thorough research on the subject. Since larger language varieties contribute more training signal during metric development, studies have observed that major linguistic variants tend to be rated more favorably than minor linguistic variants, potentially leading to biased evaluations [16].

Furthermore, analyses of neural metric performance on non-English language pairs remain limited. As a result, the superiority of neural metrics cannot be indiscriminately generalized to all language combinations [23], with some evidence suggesting that performance may degrade when English is excluded from the evaluation [24].

Among the major limitations highlighted in the literature is the lack of interpretability inherent to many neural evaluation metrics, largely due to their opaque scoring mechanisms. Their black-box nature hinders an assessment of which metrics are best suited for capturing specific linguistic phenomena and complicates the selection of appropriate metrics for targeted evaluation tasks [25]. In response, recent research has increasingly emphasized evaluation methodologies grounded in human error annotations – particularly those following the MQM (Multidimensional Quality Metrics) framework – which offer fine-grained information on translation quality [12]. These span-level annotations have also been leveraged as a standardized method for deriving quality scores (eliminating the need for direct human scoring in evaluation tasks) [26], and training more interpretable quality metrics.

Parallel efforts have also turned to linguistically motivated meta-evaluation test suites and controlled experiments designed to probe metric sensitivity to specific language phenomena [27, 28].

The specialized nature of the legal domain also raises concerns about the reliability of existing evaluation metrics. Zouhar et al. [29] highlight that learned metrics exhibit a performance drop when applied to out-of-domain data, largely due to their final-stage fine-tuning process. This suggests that current training data effectively optimizes metrics for specific domains but does not generalize well beyond them. As a result, extending these evaluation metrics to other domains – such as the legal domain – may lead to performance degradation compared to the base model.

4. Methodology

4.1. Problem Definition

We establish two criteria to characterize an effective metric for our use case: the first is *absolute agreement*, defined as ranking correct translations higher than incorrect ones. We also define *relative agreement*, that is the capability to rank translations containing critical mistakes lower than those with milder ones [11].

To operationalize the differentiation, we partition the dataset for analysis. *Absolute agreement* is measured on the Whole Dataset – comprising all segments available. To measure *relative agreement,* we subsample only the segments annotated with at least one mistake, the Mistake-only Dataset.

4.2. Dataset and Human Scoring

We use the MT@BZ corpus [8], a corpus of machine-translated decrees. It comprises source, reference and candidate translations in both language directions (IT→DE and DE→IT). Each segment has been manually annotated for translation errors using a custom error taxonomy. Table 1 offers a glance into the composition of the corpus for each language direction. We notice that around 60% of all segments is correct for both language directions. To gain further insight, we compute the BLEU score between reference and candidate sentences. Notably, we find that a very high number of segments labeled as correct receives a perfect BLEU score of 100, indicating exact matches with the reference translations. This outcome has also been observed by Oliver et al. [9] in similar experiments on the same data, and is attributed to the repetitive and formulaic nature of legal language, which often leads to low lexical and syntactic variability.

To measure correlation across a range of quality levels (as defined in Section 4.1) in the absence of numerical quality scores, we assign severity weights to each error type annotated in the original dataset (see Appendix A). Given the highly specialized nature of the domain, experts with competence in the South Tyrolean legal framework and German language varieties were consulted to define severity levels for each error type. These levels were established based on both linguistic adequacy and legislative drafting requirements[5]. For a detailed qualitative analysis of the corpus mistakes, refer to De Camillis and Chiocchetti [14].

[5]For example, the South Tyrolean public administration is bound by law to use the terminology that is being officially validated by a dedicated Terminology Commission (Presidential Decree No. 574/1988, Art. 6) and to adopt gender-neutral language (Provincial Law No. 51/2010). These constraints are therefore essential quality aspects when translating official documents into this minor language variety of German.

In this manner, we can lay out a hierarchy of type-of-error severity and derive a more granular quality ranking. We apply a penalty for each error in a segment, equal to the severity weight assigned to that error type, according to the Linear Raw Scoring Model presented by Lommel et al. [30]. The sum of penalties is then deducted from a total of 100 and becomes the human score. This score reflects both the presence and severity of translation errors, thereby enabling the computation of rank-based correlation indices between human judgments and automatic metric outputs.

Segments	**IT→DE**	**DE→IT**
Error-annotated	639	622
Exact matches	741	412
Other correct	129	475
Total segments	1,509	1,509

Table 1
Composition of MT@BZ dataset. *Error-annotated* segments indicate the number of translations that have been labeled as containing at least one mistake. *Exact matches* indicate the number of correct translations that are identical to the reference. *Other correct* segments indicate the number of correct translations that are different from the reference.

4.3. Setup of Selected Metrics

This section presents the evaluation metrics employed in our study, with details on the tested methods and models provided in Table 2. Following best practices for replicability as recommended by Zouhar et al. [42] for Comet-suite metrics, we include hash codes and model identifiers in the footnotes of the present section.

String-based Metrics

BLEU [31] measures modified n-gram precision with a brevity penalty. **chrF** [34] computes overlap over character-level n-grams, offering sensitivity to morphological and orthographic variation. Finally, **TER** [39] estimates the minimum number of edit operations required to transform the candidate into the reference, approximating post-editing effort.

Embedding-based Metrics

We utilize the BERTScore framework[6] [33], which uses contextual embeddings from pre-trained language models to compute semantic similarity. The framework allows for model selection. Hash identifiers have been

[6]https://github.com/Tiiiger/bert_score

Metric	Type	Source	Reference	Error span	Citation
BLEU	String-based	✗	✓	✗	[31]
BLEURT	Learned	✗	✓	✗	[32]
BERTScore	Embedding-based	✗	✓	✗	[33]
chrF	String-based	✗	✓	✗	[34]
COMET-22-DA	Learned	✓	✓	✗	[35]
COMET-Kiwi-DA	Learned	✓	✗	✗	[36]
COMET-KiwiXL-DA	Learned	✓	✗	✓	[37]
MetricX-24-Hybrid	Learned	✓	✓	✓	[38]
TER	String-based	✗	✓	✗	[39]
UNITE	Learned	✓	✓	✗	[40]
XCOMETXL-DA	Learned	✓	✓	✓	[41]

Table 2
Details about the evaluation models and methods considered in the study, in alphabetical order.

generated together with the scores and are provided in the footnotes. In our experiments, we evaluate four encoder backbones: **bert-base-multilingual**[7] (which is the default model), **roberta-large-mnli**[8], **deberta-xlarge-mnli**[9] and **bart-large-mnli**[10].

In Table 3, we report the results under their respective model denominations, matched with the aggregated F1 score. We also compute *precision* and *recall* individually to highlight asymmetric contributions to the similarity assessment, which will be commented in Section 5. *Precision* measures how many of the candidate's tokens are present in the reference, while *recall* captures how well the reference tokens are matched by the generated candidate.

Learned Metrics

We choose learned metrics trained under different input configurations.

We begin with *reference-based* metrics, which incorporate the reference translation during both training and inference. We select **COMET-22-DA**[11] [35] and **BLEURT** [32], which have been fine-tuned simply using quality scores from human annotators.

We also consider *source-based* metrics (also called Quality Estimation or QE metrics), which are trained without access to reference translations. Instead, they learn to predict human quality scores solely from the source sentence and the machine-generated output. We include both **COMET-Kiwi-DA**[12] [36] and its larger variant **COMET-KiwiXL-DA**[13] [37], which builds on the same architecture but differs in model capacity.

The *unified* approach combines both the source and the reference to exploit multi-task interaction. We assess **UNITE**[14] [40]. It jointly leverages the source and the reference as separate input streams during training, then incorporating a last layer to fuse the decomposed scores into the holistic one. We report scores for source (*src*) and reference (*ref*) decompositions.

We also include *error-span* metrics, namely **XCOMETXL-DA**[15] [41], **MetricX-24-Hybrid-Large** and its larger configuration **MetricX-24-Hybrid-XL** [38]. These metrics include a training phase based on error-span labels, according to the MQM error taxonomy. They are trained to predict error spans alongside a penalty score. XCometXL-DA is a hybrid metric that provides additional scores based on four decomposed dimensions: *src*, *ref*, *unified* approach and *MQM* annotations. The holistic score is then produced by ensembling the four sub-scores via a forward pass that establishes aggregation weights. Instead, the MetricX model suite only provides a single additional decomposed score which includes only the source in the evaluation.

Finally, we explore a variant of XCOMETXL quantized to **8 bits**[16], motivated by the hypothesis put forward in Zouhar et al. [42] that lower precision approximations of large metrics can maintain correlation with human judgments while significantly reducing inference costs.

4.4. Meta-Evaluation

In Table 3, we report Accuracy *(Acc)* [11], a measure computed through pairwise comparisons across the test set. It quantifies the proportion of pairs for which the

[7] bert-base-multilingual-cased_L9_no-idf_version=0.3.12(hug_trans=4.46.2)_fast-tokenizer
[8] roberta-large-mnli_L19_no-idf_version=0.3.12(hug_trans=4.51.3)
[9] microsoft/deberta-xlarge-mnli_L40_no-idf_version=0.3.12(hug_trans=4.51.3)
[10] facebook/bart-large-mnli_L11_no-idf_version=0.3.12(hug_trans=4.51.3)
[11] Python3.8.10|Comet2.2.2|fp32|Unbabel/wmt22-comet-da|1
[12] Python3.8.10|Comet2.2.2|fp32|Unbabel/wmt22-cometkiwi-da|1
[13] Python3.8.10|Comet2.2.2|fp32|Unbabel/wmt22-cometkiwiXL-da|1
[14] Python3.8.10|Comet2.2.2|fp32|Unbabel/unite-mup|1
[15] Python3.8.10|Comet2.2.3|fp32|Unbabel/XCOMET-XL|1
[16] Python3.8.10|Comet2.2.2|qint8|Unbabel/XCOMET-XL|1

evaluation metric produces the same relative ordering as the human gold standard (concordant), versus those where the ordering is incorrect (discordant). We follow Deutsch et al. [43] by using a variant of Accuracy adjusted for tie calibration by artificially creating ties from continuous scores. This procedure is needed in the light of the high number of rank ties stemming from human score fabrication. The *Acc* value ranges from 0 to 1.

We also adopt Spearman's correlation (*Rho*) (ranging from -1 to 1). It offers robustness to outliers and allows to capture rank-based monotonic relationships even across the markedly different score distributions observed in the metrics evaluated [3].

We decide not to use Pearson's correlation because it assumes a linear relationship between the distributions of the two score groups [43]. The proportional severity weights we assign to different error types are not expected to be linearly replicated by metric outputs.

5. Results

We apply meta-evaluation measures on both the Whole Dataset and the Mistake-only Dataset. This addresses the need to adequately test the metrics on the two criteria that have been established when defining the problem in Section 4.1: *absolute* and *relative* agreement. In Table 3, results are accordingly structured under two main sections, which separately report metric performance under each evaluation criterion. For metrics that generate holistic scores by aggregating subscores algorithmically, we report the holistic score in bold, while single decomposed scores are provided in regular font.

Metric paradigm performance varies across quality ranges. Our results reveal a widely different performance pattern across metric paradigms when evaluated on the Whole Dataset versus the Mistake-only Dataset. Surprisingly, both string-based and embedding-based metrics outperform learned metrics when evaluated on the Whole Dataset. We explain this with the argument that string-based metrics – being rule-based – can reliably detect and reward the high sample of exact matches with the reference. Embedding-based metrics also benefit from their ability to capture lexical overlap at a subword or token level, recognising meaning even when the wording differs. We attribute the underperformance of learned metrics primarily to the inherent nature of their regression-based scoring. Unlike rule-based metrics that produce deterministic outputs, learned metrics rely on regression functions that approximate scores based on distributional patterns in the training data. This can result in unexpected behavior – for instance, candidate translations identical to the reference may not receive the maximum score, or scores may fall outside the valid range of 0 to 1 as in Comet models (requiring post-hoc clipping). This behavior is consistent with prior observations about learned metrics' underperformance on high-quality translations, as noted by Agrawal et al. [44].

However, the trend reverses in the Mistake-only Dataset: here, when including the reference, learned metrics consistently outperform other metric types, regardless of the statistical measure used. This suggests that their modeling power becomes more effective in lower-quality bands, where surface-level matches are less common and mistakes have to be properly identified and penalized. Despite this regained advantage in the Mistake-only setting, the overperformance margins of most learned metrics remain tight and agreement levels insufficient for a reliable quality evaluation. This suggests that there is still room for improvement – especially as far as smaller-size metrics are concerned.

Mind the reference. Disaggregating the performance of learned metrics by input type offers valuable insights into which linguistic resources most effectively contribute to accurate evaluation. Considering the Mistake-only Dataset, *reference-based* scores surpass both *source-based* and *error-span* counterparts for COMET, UNITE and MetricX families. Interestingly, for metrics built on the *unified* approach (such as UNITE and XCOMETXL), the inclusion of both source and reference appears beneficial. While the reference remains the primary driver of correlation, incorporating the source provides a modest boost to overall score agreement. This suggests that *unified* models, which incorporate additional layers to weigh and integrate information streams from both inputs into the holistic score, may be better suited to capture certain error types that are only apparent when the source is considered.

In general, while *source-based* metrics trail behind other learned metric types, they can outperform embedding-based metrics counting on reference translations, especially if we consider models with larger capacity (MetricX-24-XL-QE, COMET-KiwiXL-DA and XCOMETXL-src).

Error-span metrics are misaligned. We assess the usefulness of error-span annotations in comparison to other linguistic signals. XCOMETXL-DA-mqm is the only available decomposed score based exclusively on MQM error span identification. Considering the Mistake-only Dataset, we observe a drop compared to related subscores of the same metric as well as to the smaller configuration of the same metric (COMET-22-DA). This failure may be attributable to a misalignment between the MQM

		WHOLE DATASET				MISTAKE-ONLY DATASET			
Type	**Metric**	**IT→DE**		**DE→IT**		**IT→DE**		**DE→IT**	
		Acc	**Rho**	**Acc**	**Rho**	**Acc**	**Rho**	**Acc**	**Rho**
String-based	**BLEU**	0.777	0.768	0.668	0.694	0.509	0.157	0.552	0.250
String-based	**chrF**	0.775	0.761	0.717	0.688	0.525	0.217	0.529	0.179
String-based	**TER**	0.776	0.771	0.720	0.703	0.505	0.175	0.522	0.174
Embedding	**bert-base-multilingual**	0.781	0.724	0.715	0.724	0.527	0.267	0.549	0.289
Embedding	**bart-large-mnli**	0.780	0.773	0.755	0.728	0.529	0.254	0.530	0.213
Embedding	**deberta-xlarge-mnli**	0.779	0.771	0.739	0.728	0.526	0.258	0.533	0.227
Embedding	**roberta-large-mnli**	0.771	0.758	0.760	0.738	0.524	0.252	0.524	0.221
Learned	**BLEURT**	0.706	0.686	0.660	0.512	0.488	0.123	0.551	0.255
Learned	**COMET-22-DA**	0.670	0.680	0.665	0.686	0.565	0.375	0.566	0.332
Learned	**COMET-Kiwi-DA**	0.441	0.242	0.401	0.178	0.474	0.116	0.520	0.219
Learned	**COMET-KiwiXL-DA**	0.411	0.221	0.403	0.177	0.540	0.293	0.545	0.250
Learned	**MetricX-24-Large**	0.615	0.582	0.586	0.548	0.538	0.272	0.592	0.363
Learned	MetricX-24-Large (src)	0.418	0.234	0.405	0.175	0.495	0.143	0.523	0.168
Learned	**MetricX-24-XL**	0.607	0.612	0.586	0.535	0.536	0.272	0.612	0.419
Learned	MetricX-24-XL (src)	0.463	0.554	0.409	0.189	0.494	0.143	0.554	0.254
Learned	**UNITE**	0.711	0.691	0.654	0.644	0.527	0.240	0.558	0.265
Learned	UNITE (src)	0.407	0.194	0.398	0.187	0.475	0.091	0.518	0.152
Learned	UNITE (ref)	0.745	0.717	0.660	0.664	0.529	0.248	0.563	0.285
Learned	**XCOMETXL-DA**	0.508	0.426	0.547	0.459	0.582	0.436	0.616	0.496
Learned	XCOMETXL-DA (src)	0.406	0.177	0.415	0.190	0.551	0.343	0.595	0.408
Learned	XCOMETXL-DA (ref)	0.544	0.464	0.553	0.489	0.579	0.434	0.618	0.491
Learned	XCOMETXL-DA (MQM)	0.530	0.455	0.586	0.519	0.547	0.425	0.541	0.399
Learned	XCOMETXL-DA (unified)	0.507	0.390	0.534	0.444	0.577	0.505	0.627	0.505
Learned	XCOMETXL-DA (8bit)	0.449	0.362	0.503	0.423	0.589	0.447	0.613	0.449

Table 3
The **Metric** columns shows the name of the metric: metrics in bold represent holistic scores, while metrics in regular font show decomposed scores. The **Whole Dataset** section denotes results obtained on all segments available in the dataset. The **Mistake-only Dataset** section indicates the results obtained onto a subset of the whole dataset comprising only segments containing at least one mistake. *Acc* denotes the tie-adjusted Accuracy measure, while *Rho* stands for the Spearman's correlation measure. The strongest statistical correlation for every column is underlined.

annotation framework used for training such metrics and our custom error taxonomy used for evaluation. Striving for consistency over error label criteria across training and evaluation is thus fundamental for fair assessment.

Looking at Whole Dataset, we likewise highlight that *error-span* metrics (MetricX and XCOMETXL) are surpassed by learned metrics that are optimized only for direct scalar prediction of sentence-level quality, such as COMET-22-DA, BLEURT and UNITE. As the training objective of *error-span* metrics is to regress over error annotations to estimate penalty weights accordingly, they may show a proneness for over-correction even in high-quality segments.

Precision or Recall? In Appendix B, we collect decomposed subscores for embedding-based metrics: *recall* and *precision*. We notice that *recall* tends to correlate more strongly with human judgments than the holistic score and the *precision* subscore. This trend may corroborate the importance of the reference translation: gauging how much of semantic and syntactic information contained in the reference transfers to the candidate may generally serve as a predictor of legal text quality as conceived of by expert evaluators. Yet, the negligible edge in the correlation measure is neither strong nor consistent enough to draw definitive conclusions. An informed interpretation of the results would require a qualitative analysis on the amount of semantic explicitation commonly expressed in the legal texts of both languages.

Minor varieties remain penalised. Focusing on the target language, we observe that correlations on the Mistake-only Dataset are generally higher for Italian than for German. This result is noteworthy given that German benefits from a larger pool of training data due to the fact that it is a more regularly featured language in WMT shared tasks, which contribute most of metrics training

data. We posit that this discrepancy supports the argument that generic models tend to embed biases toward dominant language varieties. In the case of German, it is likely that the datasets used to train evaluation metrics predominantly feature standard varieties such as those used in Germany and at the EU level.

Moreover, we caution against drawing conclusions based on the Whole Dataset, where Italian-to-German translations include nearly twice as many full matches between reference and candidate as the reverse direction. This makes the datasets not comparable to each other, inflating metric performance and simplifying evaluation for German as the target language.

Size matters. When comparing learned metrics of increasing model size on the Mistake-only Dataset, we observe a general trend where scaling up benefits evaluation performance. This is evident in the case of COMET-Kiwi, where the XL variant consistently outperforms its smaller counterpart, and for *reference-based* scores of XCOMETXL-DA-ref, which shows stronger results compared to COMET-22-DA. A more nuanced picture emerges with MetricX, where the XL versions outperform the Large models only in evaluations into Italian, suggesting that scaling effects may vary across language directions, presumably due to the language variety provenance of additional data.

The quantized version of XCOMETXL-DA, though slightly lowering correlation measures compared to its full-precision counterpart, still outperforms all other metrics, which confirms previous findings that quantization can be a viable strategy for reducing computational costs.

6. Conclusions

As an indication for future metric development, we conclude that reference translations are most crucial for enhancing evaluation reliability, while source sentences may contribute marginally but are not essential. We advise against embarking on the effort of error-span annotation of large corpora with the aim of training new metrics: it has notable human and resource costs but results offer no evidence that they determine commensurate metric improvements. Instead, targeted extensions of the existing MT@BZ dataset may provide more cost-effective support for evaluation purposes.

Given the underperformance of metrics when evaluating South Tyrolean German as a target language, future metric adaptation would likely benefit from applying continued pre-training to generic encoder models on South Tyrolean German data. This would provide a more suitable backbone for further fine-tuning learned metrics. To this end, efforts should be made to compile legal text corpora in South Tyrolean German and including relevant terminology.

Also, we recommend exploring training strategies that integrate the strengths of embedding-based and learned metrics, with the goal of developing evaluation systems that perform robustly across the full quality spectrum of machine translation output.

From a broader perspective, we suggest that metric selection in natural language generation tasks should be guided by a clear definition of the evaluation objective and the nature of the task. Learned metrics are more effective when the task involves detecting and weighing complex linguistic phenomena that may surface in diverse forms – such as in summarization or question-answering tasks. In such cases, the fine-tuning and validation of a custom metric may be a further convenient step. Conversely, more naive evaluation methods like the string-based ones are often appropriate when low variance from a reference is expected, such as in the presence of named entities. As our findings show, the two metric paradigms can even be complementary: embedding- and string-based metrics are well-suited for evaluating accuracy-related aspects, while learned metrics can offer global insight into the overall fluency of the generated text and meaning preservation.

References

[1] M. Zampieri, P. Nakov, Y. Scherrer, Natural language processing for similar languages, varieties, and dialects: A survey, Natural Language Engineering 26 (2020) 595–612. doi:10.1017/S1351324920000492.

[2] M. M. I. Alam, S. Ahmadi, A. Anastasopoulos, CODET: A benchmark for contrastive dialectal evaluation of machine translation, in: Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1790–1859. URL: https://aclanthology.org/2024.findings-eacl.125/.

[3] J. Wang, D. I. Adelani, S. Agrawal, M. Masiak, R. Rei, E. Briakou, M. Carpuat, et al., AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages, in: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 5997–6023. URL: https://aclanthology.org/2024.naacl-long.334/. doi:10.18653/v1/2024.naacl-long.334.

[4] J. Falcão, C. Borg, N. Aranberri, K. Abela, COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque,

in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 3553–3565. URL: https://aclanthology.org/2024.lrec-main.315/.

[5] A. Magueresse, V. Carles, E. Heetderks, Low-resource languages: A review of past work and future challenges, 2020. URL: https://arxiv.org/abs/2006.07264. arXiv:2006.07264.

[6] R. Knowles, S. Larkin, C.-K. Lo, MSLC24: Further challenges for metrics on a wide landscape of translation quality, in: Proceedings of the Ninth Conference on Machine Translation, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 475–491. URL: https://aclanthology.org/2024.wmt-1.34/. doi:10.18653/v1/2024.wmt-1.34.

[7] V. Dewangan, B. R. S, G. Suri, R. Sonavane, When every token counts: Optimal segmentation for low-resource language models, in: Proceedings of the First Workshop on Language Models for Low-Resource Languages, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2025, pp. 294–308. URL: https://aclanthology.org/2025.loreslm-1.24/.

[8] F. De Camillis, E. W. Stemle, E. Chiocchetti, F. Fernicola, The MT@BZ corpus: machine translation & legal language, in: Proceedings of the 24th Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Tampere, Finland, 2023, pp. 171–180. URL: https://aclanthology.org/2023.eamt-1.17/.

[9] A. Oliver, S. Alvarez-Vidal, E. Stemle, E. Chiocchetti, Training an NMT system for legal texts of a low-resource language variety south tyrolean German - Italian, in: Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), European Association for Machine Translation (EAMT), Sheffield, UK, 2024, pp. 573–579. URL: https://aclanthology.org/2024.eamt-1.47/.

[10] S. Perrella, L. Proietti, A. Scirè, E. Barba, R. Navigli, Guardians of the machine translation meta-evaluation: Sentinel metrics fall in!, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 16216–16244. URL: https://aclanthology.org/2024.acl-long.856/. doi:10.18653/v1/2024.acl-long.856.

[11] T. Kocmi, C. Federmann, R. Grundkiewicz, M. Junczys-Dowmunt, H. Matsushita, A. Menezes, To ship or not to ship: An extensive evaluation of automatic metrics for machine translation, in: Proceedings of the Sixth Conference on Machine Translation, Association for Computational Linguistics, Online, 2021, pp. 478–494. URL: https://aclanthology.org/2021.wmt-1.57/.

[12] M. Freitag, G. Foster, D. Grangier, V. Ratnakar, Q. Tan, W. Macherey, Experts, errors, and context: A large-scale study of human evaluation for machine translation, Transactions of the Association for Computational Linguistics 9 (2021) 1460–1474. URL: https://aclanthology.org/2021.tacl-1.87/. doi:10.1162/tacl_a_00437.

[13] F. D. Camillis, La traduzione non professionale nelle istituzioni pubbliche dei territori di lingua minoritaria: il caso di studio dell'amministrazione della Provincia autonoma di Bolzano, Ph.D. thesis, alma, 2021. URL: https://amsdottorato.unibo.it/id/eprint/9695/.

[14] F. De Camillis, E. Chiocchetti, Machine-translating legal language: error analysis on an italian-german corpus of decrees, Terminology science & research 27 (2024) 1–27. URL: https://journal-eaft-aet.net/index.php/tsr/article/view/8304/7492.

[15] J. O. Alabi, D. I. Adelani, M. Mosbach, D. Klakow, Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 4336–4349. URL: https://aclanthology.org/2022.coling-1.382/.

[16] J. Sun, T. Sellam, E. Clark, T. Vu, T. Dozat, D. Garrette, A. Siddhant, J. Eisenstein, S. Gehrmann, Dialect-robust evaluation of generated text, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 6010–6028. URL: https://aclanthology.org/2023.acl-long.331/. doi:10.18653/v1/2023.acl-long.331.

[17] C. Amrhein, N. Moghe, L. Guillou, ACES: Translation accuracy challenge sets for evaluating machine translation metrics, in: Proceedings of the Seventh Conference on Machine Translation (WMT), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 479–513. URL: https://aclanthology.org/2022.wmt-1.44/.

[18] Y. Yan, T. Wang, C. Zhao, S. Huang, J. Chen, M. Wang, BLEURT has universal translations: An analysis of automatic metrics by minimum risk training, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5428–5443. URL: https://aclanthology.org/2023.acl-long.297/. doi:10.18653/v1/2023.acl-long.297.

[19] P. Fernandes, A. Farinhas, R. Rei, J. G. C. de Souza,

P. Ogayo, G. Neubig, A. Martins, 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, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1396–1412. URL: https://aclanthology.org/2022.naacl-main.100/. doi:10.18653/v1/2022.naacl-main.100.

[20] G. Kovacs, D. Deutsch, M. Freitag, Mitigating metric bias in minimum Bayes risk decoding, in: Proceedings of the Ninth Conference on Machine Translation, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 1063–1094. URL: https://aclanthology.org/2024.wmt-1.109/. doi:10.18653/v1/2024.wmt-1.109.

[21] J. Pombal, N. M. Guerreiro, R. Rei, A. F. T. Martins, Adding chocolate to mint: Mitigating metric interference in machine translation, 2025. URL: https://arxiv.org/abs/2503.08327. arXiv:2503.08327.

[22] M. Freitag, N. Mathur, D. Deutsch, C.-K. Lo, E. Avramidis, R. Rei, B. Thompson, F. Blain, T. Kocmi, J. Wang, D. I. Adelani, M. Buchicchio, C. Zerva, A. Lavie, Are LLMs breaking MT metrics? results of the WMT24 metrics shared task, in: Proceedings of the Ninth Conference on Machine Translation, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 47–81. URL: https://aclanthology.org/2024.wmt-1.2/. doi:10.18653/v1/2024.wmt-1.2.

[23] N. Moghe, T. Sherborne, M. Steedman, A. Birch, Extrinsic evaluation of machine translation metrics, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 13060–13078. URL: https://aclanthology.org/2023.acl-long.730/. doi:10.18653/v1/2023.acl-long.730.

[24] S. Agrawal, A. Farajian, P. Fernandes, R. Rei, A. F. T. Martins, Assessing the role of context in chat translation evaluation: Is context helpful and under what conditions?, Transactions of the Association for Computational Linguistics 12 (2024) 1250–1267. URL: https://aclanthology.org/2024.tacl-1.69/. doi:10.1162/tacl_a_00700.

[25] R. Rei, N. M. Guerreiro, M. Treviso, L. Coheur, A. Lavie, A. Martins, 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), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1089–1105. URL: https://aclanthology.org/2023.acl-short.94/. doi:10.18653/v1/2023.acl-short.94.

[26] M. Freitag, N. Mathur, C.-k. Lo, E. Avramidis, R. Rei, B. Thompson, T. Kocmi, F. Blain, D. Deutsch, C. Stewart, C. Zerva, S. Castilho, A. Lavie, G. Foster, Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent, in: Proceedings of the Eighth Conference on Machine Translation, Association for Computational Linguistics, Singapore, 2023, pp. 578–628. URL: https://aclanthology.org/2023.wmt-1.51/. doi:10.18653/v1/2023.wmt-1.51.

[27] N. Moghe, A. Fazla, C. Amrhein, T. Kocmi, M. Steedman, A. Birch, R. Sennrich, L. Guillou, Machine translation meta evaluation through translation accuracy challenge sets, Computational Linguistics 51 (2025) 73–137. URL: https://aclanthology.org/2025.cl-1.4/. doi:10.1162/coli_a_00537.

[28] E. Avramidis, S. Manakhimova, V. Macketanz, S. Möller, Machine translation metrics are better in evaluating linguistic errors on LLMs than on encoder-decoder systems, in: Proceedings of the Ninth Conference on Machine Translation, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 517–528. URL: https://aclanthology.org/2024.wmt-1.37/. doi:10.18653/v1/2024.wmt-1.37.

[29] V. Zouhar, S. Ding, A. Currey, T. Badeka, J. Wang, B. Thompson, Fine-tuned machine translation metrics struggle in unseen domains, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 488–500. URL: https://aclanthology.org/2024.acl-short.45/. doi:10.18653/v1/2024.acl-short.45.

[30] A. Lommel, S. Gladkoff, A. Melby, S. E. Wright, I. Strandvik, K. Gasova, A. Vaasa, A. Benzo, R. Marazzato Sparano, M. Foresi, J. Innis, L. Han, G. Nenadic, The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control, in: Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations), Association for Machine Translation in the Americas, Chicago, USA, 2024, pp. 75–94. URL: https://aclanthology.org/2024.amta-presentations.6/.

[31] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.

[32] T. Sellam, D. Das, A. Parikh, BLEURT: Learning

robust metrics for text generation, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7881–7892. URL: https://aclanthology.org/2020.acl-main.704/. doi:10.18653/v1/2020.acl-main.704.

[33] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, 2020. URL: https://arxiv.org/abs/1904.09675. arXiv:1904.09675.

[34] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049/. doi:10.18653/v1/W15-3049.

[35] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 2685–2702. URL: https://aclanthology.org/2020.emnlp-main.213/. doi:10.18653/v1/2020.emnlp-main.213.

[36] R. Rei, M. Treviso, N. M. Guerreiro, C. Zerva, A. C. Farinha, C. Maroti, J. G. C. de Souza, T. Glushkova, D. Alves, L. Coheur, A. Lavie, A. F. T. Martins, CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task, in: Proceedings of the Seventh Conference on Machine Translation (WMT), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 634–645. URL: https://aclanthology.org/2022.wmt-1.60/.

[37] R. Rei, N. M. Guerreiro, J. Pombal, D. van Stigt, M. Treviso, L. Coheur, J. G. C. de Souza, A. Martins, Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task, in: Proceedings of the Eighth Conference on Machine Translation, Association for Computational Linguistics, Singapore, 2023, pp. 841–848. URL: https://aclanthology.org/2023.wmt-1.73/. doi:10.18653/v1/2023.wmt-1.73.

[38] J. Juraska, D. Deutsch, M. Finkelstein, M. Freitag, MetricX-24: The Google submission to the WMT 2024 metrics shared task, in: Proceedings of the Ninth Conference on Machine Translation, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 492–504. URL: https://aclanthology.org/2024.wmt-1.35/. doi:10.18653/v1/2024.wmt-1.35.

[39] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA, 2006, pp. 223–231. URL: https://aclanthology.org/2006.amta-papers.25/.

[40] Y. Wan, D. Liu, B. Yang, H. Zhang, B. Chen, D. Wong, L. Chao, UniTE: Unified translation evaluation, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8117–8127. URL: https://aclanthology.org/2022.acl-long.558/. doi:10.18653/v1/2022.acl-long.558.

[41] N. M. Guerreiro, R. Rei, D. v. Stigt, L. Coheur, P. Colombo, A. F. T. Martins, xcomet: Transparent machine translation evaluation through fine-grained error detection, Transactions of the Association for Computational Linguistics 12 (2024) 979–995. URL: https://aclanthology.org/2024.tacl-1.54/. doi:10.1162/tacl_a_00683.

[42] V. Zouhar, P. Chen, T. K. Lam, N. Moghe, B. Haddow, Pitfalls and outlooks in using COMET, in: Proceedings of the Ninth Conference on Machine Translation, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 1272–1288. URL: https://aclanthology.org/2024.wmt-1.121/. doi:10.18653/v1/2024.wmt-1.121.

[43] D. Deutsch, G. Foster, M. Freitag, 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, Association for Computational Linguistics, Singapore, 2023, pp. 12914–12929. URL: https://aclanthology.org/2023.emnlp-main.798/. doi:10.18653/v1/2023.emnlp-main.798.

[44] S. Agrawal, A. Farinhas, R. Rei, A. Martins, Can automatic metrics assess high-quality translations?, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 14491–14502. URL: https://aclanthology.org/2024.emnlp-main.802/. doi:10.18653/v1/2024.emnlp-main.802.

A. Custom error weights

Type of Error	Penalty Weight
Accuracy errors	
Mistranslation:	
Multiword expressions	20
Part of Speech	20
Word Sense Disambiguation	25
Partial	20
Semantically Unrelated	20
Addition	15
Omission	15
Untranslated	20
Mechanical	15
Bilingual terminology	25
Source error	15
Fluency errors	
Grammar:	
Multiword syntax	15
Word form	15
Word order	15
Extra words	15
Missing words	20
Lexicon:	
Lexical choice	15
Non-existing or Foreign Word	20
Orthography:	
Spelling	12
Punctuation	12
Capitalization	12
Gender	5
Inconsistency	5
Coherence	5
Multiple fluency errors	10
Other	5

Table 4
The left-hand column lists the error types defined in the custom annotation scheme, while the right-hand column shows the corresponding penalty weights applied to the segment's quality score when each error type is present. The SCATE taxonomy differentiates between fluency and accuracy errors. Some error types are grouped under higher-ranking categories (shown in underlined font), which serve only as structural labels and do not carry additional penalty weights.

B. Subscores of Embedding-based Metrics

	WHOLE DATASET		MISTAKE-ONLY DATASET	
Model	**IT→DE**	**DE→IT**	**IT→DE**	**DE→IT**
bert-base-multilingual	0.781	0.715	0.527	0.549
precision	0.778	0.721	0.517	0.546
recall	0.781	0.699	0.530	0.554
bart-large-mnli	0.780	0.755	0.529	0.530
precision	0.778	0.757	0.521	0.540
recall	0.780	0.738	0.530	0.544
deberta-xlarge-mnli	0.779	0.739	0.526	0.533
precision	0.777	0.741	0.520	0.536
recall	0.782	0.733	0.533	0.539
roberta-large-mnli	0.771	0.760	0.524	0.524
precision	0.766	0.757	0.518	0.534
recall	0.775	0.754	0.526	0.536

Table 5
Accuracy (*Acc*) correlation for the decomposed scores of the embedding-based metrics. The name of the model is in bold font, while *precision* and *recall* decompositions are written in regular font.

Declaration on Generative AI

During the preparation of this work, the author(s) did not use any generative AI tools or services.