Who Watches the Watchmen?
Humans Disagree With Translation Metrics on Unseen Domains

Finn Schmidt **Jan Philip Wahle** **Terry Ruas** **Bela Gipp**
University of Göttingen, Germany
finn.schmidt@uni-goettingen.de

Abstract

Automatic evaluation metrics are central to the development of machine translation systems, yet their robustness under domain shift remains unclear. Most metrics are developed on the Workshop on Machine Translation (WMT) benchmarks, raising concerns about their robustness to unseen domains. Prior studies that analyze unseen domains vary translation systems, annotators, or evaluation conditions, confounding domain effects with human annotation noise. To address these biases, we introduce a systematic multi-annotator **C**ross-**D**omain **E**rror-**S**pan-**A**nnotation dataset (CD-ESA), comprising 18.8k human error span annotations across three language pairs, where we fix annotators within each language pair and evaluate translations of the same six translation systems across one seen news domain and two unseen technical domains. Using this dataset, we first find that automatic metrics appear surprisingly robust to domain-shifts at the segment level (up to 0.69 agreement), but this robustness largely disappears once we account for human label variation. Averaging annotations increases inter-annotator agreement by up to +0.11. Metrics struggle on the unseen chemical domain compared to humans (inter-annotator agreement of 0.78–0.83 vs. 0.96). We recommend comparing metric–human agreement against inter-annotator agreement, rather than comparing raw metric–human agreement alone, when evaluating across different domains.

1 Introduction

Machine translation aims to break barriers between different cultures, supporting humanity's ability to collaborate and share knowledge (Steigerwald et al., 2022). To quantify which translation methods are "good", we need robust quality estimation (QE) metrics that can assess translation quality based solely on the source and candidate translation,

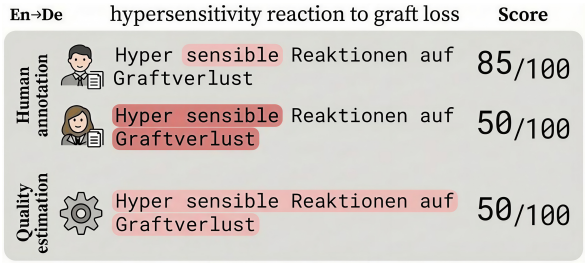


Figure 1: Metrics may show unexpected behavior on unseen domains, reflecting domain familiarity over translation quality, while variability in human judgments (e.g., overlooking errors) can mask these effects.

as reference translations are often unavailable when exploring new domains (Mathur et al., 2020). Those QE metrics enable large-scale experimentation and frequent model comparisons that would be too costly with human annotation (Chaganty et al., 2018). Although many studies rely on evaluation metrics as the *watchmen* for progress in the field, there are valid concerns whether these metrics accurately quantify progress as intended, especially in low-resource settings (Kocmi et al., 2021).

Most metric development and validation is centralized around the WMT shared tasks (Lavie et al., 2025). The WMT shared tasks provide a large, annually refreshed human-evaluated benchmark of bilingual text drawn from multiple high-resource domains such as news, literary, social, and spoken sources. This has raised concerns that metric performance degrades under conditions outside this distribution, for example on medical data.

Few studies have directly examined QE metric robustness on unseen domains by collecting new human annotations tailored to out-of-domain evaluation (Zhang et al., 2025; Zouhar et al., 2024). However, comparing domains between studies is difficult and assessments are often done by different human evaluators, raising two

Findings of the Association for Computational Linguistics: ACL 2026, pages 22822–22841

methodological gaps to motivate this study. First, human judgments are known to be inconsistent (Belz et al., 2023), especially at the segment level (Plank, 2022). Individual ratings contain noise that can act as a bottleneck for reliable metric evaluation (illustrated in Figure 1). Second, studies that change annotator pools, translation methods, and annotation conditions across datasets make it challenging to compare results and attribute performance differences solely to domain shift (Zouhar et al., 2024). Therefore, raw agreement or correlation scores are difficult to interpret, as they confound metric behavior with domain-specific annotation difficulty (Agrawal et al., 2024). We argue that variation in human label consistency across domains and annotator pools makes machine translation scores difficult to compare. In effect, we need better ways for assessing the QE metrics in out-of-domain settings, this *watching the watchmen*.

In this study, we introduce the **C**ross-**D**omain **E**rror-**S**pan-**A**nnotation dataset (CD-ESA) dataset with 18.8k human error span annotations from 2 or 4 annotators per language pair on English-Chinese (en-zh), English–German (en-de), and English-Korean (en-ko) translations across three domains: WMT23 news, public medical documents (Emea), and proprietary pharmaceutical and chemical data (PharmaChem). We ensure fair cross-domain comparison by evaluating translations of the same MT systems and relying on the same annotators per language pair across all domains. Each segment is annotated by multiple human annotators for consistent measurement of inter-annotator agreement and fair comparison of annotation difficulty across domains. Since human labels can vary substantially (Plank, 2022; Sarti et al., 2025), we examine how averaging human annotations reduces this variability. We also propose a grouping strategy that combines n segments, reducing annotation costs and inconsistency by requiring fewer human judgments. To avoid data contamination, particularly for LLM-as-a-judge approaches (Kocmi and Federmann, 2023b), we derive our unseen-domain dataset from proprietary company data of shift worker log entires with technical terms and specific jargon.

We first find that QE metrics appear surprisingly robust at the segment level on unseen domains (up to 0.69 agreement), often within 0.04 of inter-annotator agreement, but this robustness largely disappears once we account for human label inconsistency. Averaging annotations increases inter-annotator agreement by up to +0.11. Metrics struggle in the unseen PharmaChem domain, inter-annotator agreement reaches 0.96, while metrics often achieve only 0.78–0.87. Metrics underestimate translation quality by up to 0.37 normalized score points and sometimes overprefer open-source MT systems. These effects are strongest for fine-tuned metrics such as XCOMET_QE (Guerreiro et al., 2024), while LLM-as-a-judge approaches show greater robustness.

Our CD-ESA annotations for WMT and Emea across all three langauge pairs are publicly available.[1] PharmaChem contains proprietary data and cannot be released.

Key Contributions:

- **A Cross-Domain ESA Dataset.** We introduce **CD-ESA**, a multi-annotator **C**ross-**D**omain **E**rror **S**pan **A**nnotation dataset for three (unseen) domains under consistent evaluation conditions.
- **Disentangling metric robustness from human label inconsistency.** To reduce human label inconsistency, we propose and evaluate a cost-effective alternative that groups n segments into a single evaluation unit.
- **Analysis of metric weaknesses under domain shift.** We analyze the reasons why metrics fail under domain shift.

2 Related Work

Quality estimation (QE) predicts translation quality without reference translations (Specia et al., 2009), making it key for low-resource settings, where references are often not available. Most QE research follows two paradigms. First, methods fine-tune encoder-based models on human judgments, including CometKiwi (Rei et al., 2023), MetricX (Juraska et al., 2024), and xComet (Guerreiro et al., 2024). Remedy (Tan and Monz, 2025) departs from regression-based training by learning relative quality from pairwise preferences, improving performance on standard benchmarks. Second, methods uses LLM-as-a-judge approaches, such as GEMBA-MQM (Kocmi and Federmann, 2023a), which assess translation quality via prompting (Kocmi and Federmann, 2023b; Fernandes et al., 2023). While both paradigms report strong results on benchmark

[1] https://huggingface.co/datasets/FinnSchmidt/CD-ESA

data, their robustness on new domains remains underexplored.

Meta-evaluation of QE Metric development and validation are strongly shaped by the WMT Metrics Shared Tasks, where submitted metrics are ranked by agreement with newly collected human judgments (Freitag et al., 2022b, 2023, 2024; Lavie et al., 2025). These benchmarks focus on a limited set of high-resource domains (e.g., news, literary text, reviews) and language pairs (e.g., en–de, en–zh). Over time, WMT has produced large human-annotated datasets using standardized evaluation protocols, such as Direct Assessment (DA) (Graham et al., 2013), Multidimensional Quality Metrics (MQM) (Freitag et al., 2021), and ESA (Kocmi et al., 2024). As a result, most state-of-the-art metrics train on WMT-derived judgments and achieve near-human agreement on WMT test sets (Proietti et al., 2025). We argue that such results show adaptation to WMT-style distributions. Our work evaluates metrics under matched conditions but outside this distribution.

Metrics on unseen domains Only few studies evaluate metrics using newly collected annotations on unseen domains. Zouhar et al. (2024) report degraded metric performance out of domain, though updated analyses show sensitivity to label imbalance (Section A). Further, they assume BLEU as a domain-invariant reference, despite evidence of domain-dependent performance (Section A). Moreover, prior studies vary MT systems, annotators, or annotation protocols across domains, confounding domain effects with human variability. Recent work on QE in the literary domain (Zhang et al., 2025) focuses on human evaluation reliability rather than cross-domain metric robustness.

Prior work highlights challenges in evaluating metrics on unseen domains but typically varies annotators, MT systems, or evaluation setups across datasets, making it difficult to isolate true domain effects from annotation noise. This lack of controlled cross-domain evaluation leaves it unclear whether observed performance drops reflect genuine metric weaknesses or differences in human labeling consistency.

3 Methodology

We create the CD-ESA dataset to evaluate MT metrics under domain shift by translating data from three domains with the same MT systems, and collecting multiple human error span annotations per segment. We assess metric–human agreement at segment and system levels using the WMT25 (Lavie et al., 2025) protocols and compare it with inter-annotator agreement. This design allows us to disentangle metric robustness from human label inconsistency and to analyze systematic biases in metrics across unseen domains.

3.1 Creation & annotation of CD-ESA

To create our CD-ESA dataset, we collect English source sentences from three domains:

(i) WMT23 (in-domain) A randomly sampled subset of English news data from the metrics shared task of WMT23 consisting of well-edited journalistic text from a high-resource news domain on which most automatic metrics are developed and validated.

(ii) PharmaChem (unseen, proprietary) A proprietary industrial dataset of English log entries written by shift workers[2] in pharmaceutical and chemical plants. The data contains domain-specific terminology, abbreviations, and short, non-fluent utterances. As proprietary data, PharmaChem is contamination-free for LLM-based evaluation.

(iii) Emea (unseen, public proxy) A parallel corpus derived from documents from the European Medicines Agency (Tiedemann, 2012). The data contains medical and regulatory text with specific terminology and abbreviations. As a public proxy for PharmaChem, we retain only segments classified as similar to PharmaChem using a trained binary domain classifier. We trained a SciBERT domain classifier on an 80k balanced dataset (50% unrelated domains, 50% Medicine/Pharma and sampled PharmaChem data). We then applied it to multilingual corpora, identified Emea as the closest match, and retained sentences with classifier score > 0.5 as a public proxy (for more details see Section F).

We translate all source sentences from English into German by the same six MT systems: GPT-4o (OpenAI, 2024), DeepL (DeepL SE, 2024), Azure MS (Microsoft, 2024), X-ALMA (Xu et al., 2024), Tower-Instruct-v2 (Rei et al., 2024), and a Tower model we fine-tune on PharmaChem data. This results in a total of 1354 translations (Table 1).

To ensure comparability and avoid introducing inference-related artifacts, all open-source MT systems use beam decoding (Sutskever et al., 2014).

[2]Shift workers operate industrial production processes and document processes, incidents, and handovers across shifts.

This choice avoids metric-hacking effects of other decoding strategies, such as Minimum Bayesian Risk decoding (Freitag et al., 2022a; Tomani et al., 2024; Pombal et al., 2025).

We annotate all translations using a total of five annotators and at least two annotators per segment, enabling estimates of inter-annotator agreement across domains. All annotations follow the Error Span Annotation (ESA) protocol (Kocmi et al., 2024) and are collected using Label Studio[3].

For en-de translations, annotators are native German speakers (four female, one male), aged between 20 and 25, and employed as working students with prior experience in linguistic annotation tasks. For en-ko and en-zh translations, annotators are professional translators with native proficiency in Korean and Chinese. They are recruited through a certified translation vendor.

Table 1 summarizes dataset statistics. PharmaChem is the most challenging domain, with 2.8 errors per segment on average, compared to 1.4 on WMT23 and 0.8 on Emea. The lower error rate on Emea is partly explained by shorter segments (22.4 vs. 38 tokens for WMT23 and PharmaChem).

Following human annotation, we evaluate three categories of automatic metrics:

(i) Supervised fine-tuned (SFT) metrics, including xCOMET-XL_QE, MetricX-24-XL_QE, and COMETKIWI-23-XL. Unless stated otherwise, reported SFT metrics results correspond to the mean performance across these three metrics.

(ii) Remedy, a preference-based metric that outperforms fine-tuning on recent WMT benchmarks.

(iii) GEMBA, an **LLM-as-a-judge approach** using the GEMBA-ESA-QE prompt[4] with GPT-4o[5].

Since reference translations are often unavailable in low-resource and domain-specific settings, we focus exclusively on the **QE capabilities of these metrics**. All evaluated metrics are explicitly trained or configured for reference-free evaluation.

3.2 Meta-Evaluation

To assess automatic metrics on the CD-ESA dataset, we follow the WMT25 Metrics shared task evaluation and measure metric-human agreement at system- and segment-level.

[3] https://labelstud.io

[4] https://github.com/MicrosoftTranslator/GEMBA

[5] We use the GPT-4o API version released on 2024-12-01.

Metric	Emea	Pharma-Chem	WMT23
# Annotated segments	**3060**	**3808**	**1668**
# Sources	528	652	249
Avg. words / source	28.6	36.7	46.5
# Annotators per segment for en-de/en-ko/en-zh	**4/2/2**	**2/2/2**	**4/2/2**
Avg. ESA score	80.9	65.1	83.8
Avg. errors / seg	1.4	2.9	1.6
Minor errors [%]	58.8	42.1	73.0
Major errors [%]	41.2	57.9	27.0

Table 1: Dataset overview of our new dataset CD-ESA.

Segment-level evaluation. For segment-level evaluation, we use pairwise accuracy with tie calibration (acc_{eq}) (Deutsch et al., 2023). Specifically, acc_{eq} counts how often a metric m ranks pairs of translations of the same source segment in the same order as a human annotator h, while accounting for tied scores.

System-level evaluation. We evaluate system-level metric performance using Soft Pairwise Accuracy (SPA) (Thompson et al., 2024). SPA measures how well a metric agrees with humans on the relative ranking of MT systems while accounting for the confidence of these preferences. Computational details of acc_{eq} and SPA are in Section B.

Oracle bounds We also establish performance bounds ensuring fair comparison across domains. We measure inter-annotator agreement using the same metrics (acc_{eq}, SPA) to define an upper bound. As lower bounds, we include a dummy metric that assigns random scores and a sentinel candidate metric (Perrella et al., 2024) that has access only to the candidate translation and not the source. Thus, this metric is unable to evaluate the quality of machine-translated text properly.

Unless stated otherwise, we compute evaluation scores by treating each human annotator as the ground truth individually and averaging the resulting values. For inter-annotator agreement, we compute agreement over all possible annotator pairs and report the average.

Bias analysis and implementation. To analyze potential biases, we additionally examine scores assigned by metrics and humans across MT systems and domains directly. This allows us to detect preferences for specific systems and systematic score shifts across domains. We conduct parallel analyses of human scoring to distinguish effects caused by human label inconsistency from

metric biases.

4 Experiments

Our experiments investigate how well automatic MT evaluation metrics generalize to unseen domains and identify factors that limit their reliability under domain shift.

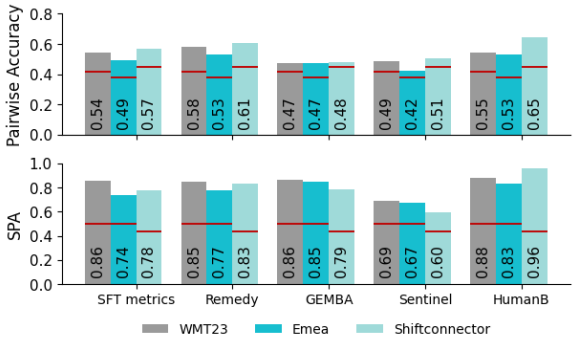
First, we assess whether current metrics exhibit degraded performance on unseen domains. We compare metric–human agreement at the segment and system levels across seen and unseen domains, following the WMT25 meta-evaluation protocol. Interpreting metric–human agreement relies on human judgments as ground truth. However, human annotations are known to exhibit substantial label inconsistency (Plank, 2022). This raises the question of whether human noise may conceal differences in metric behavior across domains. Second, we investigate the role of human label inconsistency in cross-domain evaluation. By averaging human scores and grouping multiple segments, we reduce annotation noise and obtain a cleaner evaluation signal. This allows us to test whether domain-specific effects become more apparent once human noise is mitigated. Third, we analyze the reasons metrics fail to align with human judgments on unseen domains. We focus on identifying systematic biases and domain-specific effects that explain mismatches between metric and human evaluations.

4.1 QE metrics appear robust on unseen domains

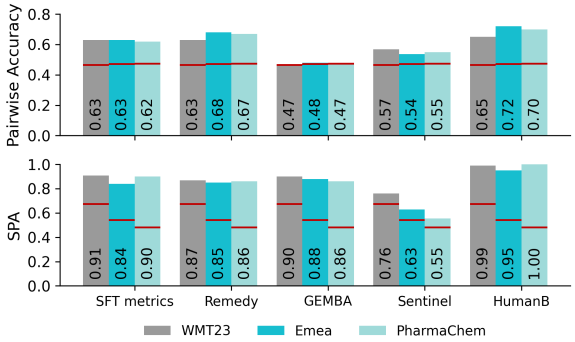
To compare QE metric performance across seen and unseen domains, we measure segment- and system-level metric–human agreement on our CD-ESA dataset across three domains: WMT23, Emea, and PharmaChem (Section 3.1) across the three language pairs en-de, en-ko, en-zh. Figure 2 reports segment-level acc_{eq} and system-level SPA.

At the segment level, metrics show no degradation on unseen domains. Across all three language pairs, SFT metrics and Remedy remain stable or even slightly improve on the unseen PharmaChem domain (0.60→0.62 and 0.62→0.69 on en-zh) compared to WMT23. GEMBA also shows consistent performance across domains (0.47-0.48 accuracy on en-de, en-ko for all domains). Overall, segment-level metric–human agreement remains stable across domains.

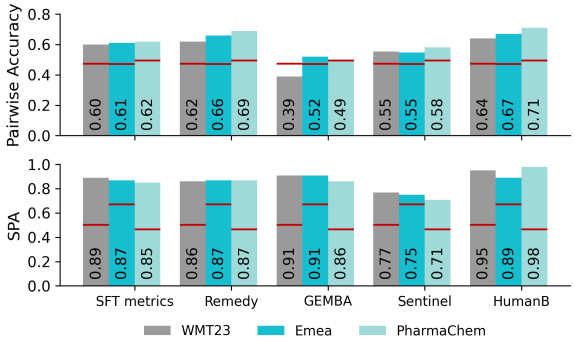
Comparison to inter-annotator agreement further



(a) English-German



(b) English-Korean



(c) English-Chinese

Figure 2: (Soft)Pairwise accuracies of metric–human and human–human (humanB) agreement on CD-ESA. The red line is the performance of a random baseline.

supports this robustness. Across all domains and language pairs, at least one metric is within approximately 0.03–0.05 of human–human agreement. For example, Remedy closely tracks human agreement across settings (e.g., differences of $\leq$0.03 on WMT23, $\leq$0.04 on Emea, and $\leq$0.04 on PharmaChem). In some cases, metric–human agreement even exceeds inter-annotator agreement (e.g., Remedy: 0.58 vs. 0.55 on WMT23 en-de), suggesting that human label inconsistency may affect reliability (Plank, 2022).

At the system level, a different pattern emerges. On WMT23, metrics achieve high agreement (typically $\sim$0.85–0.91), close to inter-annotator agreement across language pairs. On Emea,

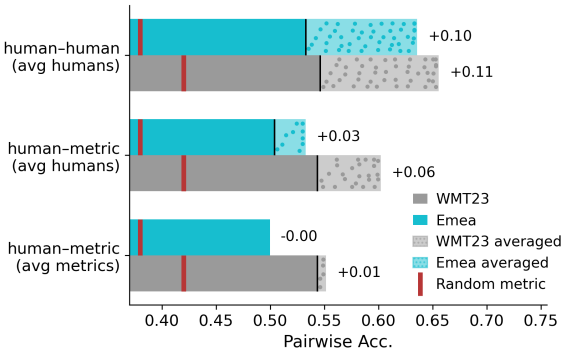


Figure 3: Pairwise accuracy gains from score aggregation on WMT23 and Emea. Bars report changes in acc_{eq} when replacing single with averaged scores.

GEMBA remains competitive (e.g., ∼0.85–0.91) and in some cases matches or slightly exceeds human agreement, while SFT metrics and Remedy tend to lag behind (e.g., ∼0.74–0.87 depending on the language pair). On PharmaChem, inter-annotator agreement is consistently high (up to ∼0.96–1.00), but metrics fall behind across all language pairs, with SPA values typically between ∼0.78-0.87. This widening gap indicates that, once human label noise is reduced, metric weaknesses on unseen domains become more apparent at the system level despite their apparent robustness at the segment level.

4.2 Averaging human annotations reveals metric weaknesses on unseen domains

Here, we want to understand whether the apparent cross-domain robustness observed in Section 4.1 reflects metric generalization or is instead driven by human label inconsistency. We evaluate the effect of averaging multiple human annotations as a noise-mitigation strategy and propose a second, cost-effective strategy that groups n segments into a single evaluation unit. As we have four instead of two annotators per segment only for en-de, we experiment only with en-de data in this experiment.

Aggregating multiple annotations. We assume that each translation has an underlying ground-truth quality score and that annotators approximate this score with independent noise (Graham et al., 2015). Under this assumption, averaging multiple annotations for the same segment should reduce noise. As a result, we expect both inter-annotator and metric–human agreement to increase as more annotations are aggregated (Sarti et al., 2025).

To enable averaging across heterogeneous scoring scales, we first z-normalize all scores for each annotator and each metric separately, following prior work (Bojar et al., 2017). We first examine inter-annotator agreement. Specifically, we compare acc_{eq} at the segment level between two individual annotators (single–single) to agreement computed between two groups of annotators, where each group's scores are averaged over two annotators (pair–pair). In both cases, we average the results over all possible single–single and pair–pair annotator combinations. Averaging human annotations leads to a consistent improvement across domains, suggesting that human label inconsistency is largely domain-independent.

As shown in the top bars of Figure 3, averaging yields consistent gains across domains: inter-annotator agreement increases from 0.55 to 0.66 on WMT23 (+0.11) and from 0.53 to 0.63 on Emea (+0.10), indicating largely domain-independent human label noise.

Next, we analyze the average metric–human agreement by replacing a single human annotation with the average of four human annotations as ground truth (middle bars in Figure 3). On WMT23, agreement increases from 0.54 to 0.60 (+0.06). On the unseen Emea domain, gains are smaller, from 0.50 to 0.53 (+0.03), despite comparable reductions in human noise. This divergence indicates that once annotation noise is reduced, metric weaknesses on unseen domains become more apparent.

Finally, we examine average metric–human agreement when averaging scores from four metrics instead of using a single metric score (bottom bars in Figure 3). In contrast to human averaging, this yields no meaningful improvement. The absence of gains indicates that current metrics exhibit correlated errors rather than independent noise, reflecting shared biases.

Grouping n segments. Since collecting multiple human annotations per segment is costly and often infeasible for large datasets, we propose an alternative noise-mitigation strategy: **grouping n segments**. Instead of averaging multiple annotations for the same segment, we group n segments translated by the same MT system and average their scores. This yields an intermediate evaluation setting between segment-level evaluation ($n = 1$) and full system-level evaluation (large n).

Figure 4 shows the results. For WMT23 with $n = 1$, the inter-annotator agreement

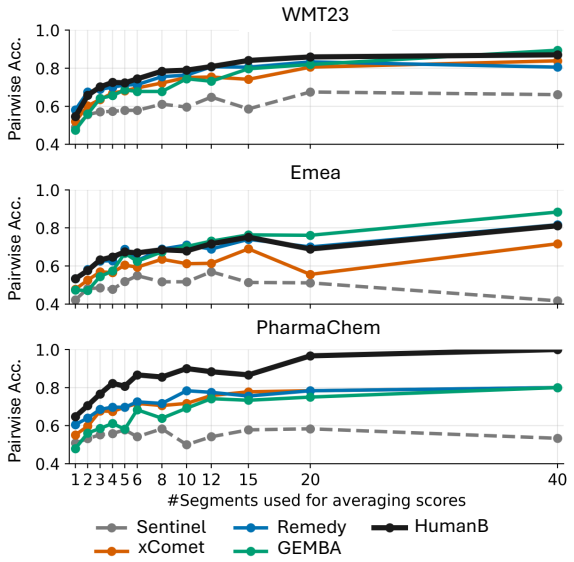


Figure 4: Pairwise accuracy for grouping segments. The x-axis shows number of segments (n), the y-axis shows pairwise accuracy. We plot inter-annotator agreement (HumanB) and metric–human agreement for several metrics with lower-bound (sentiel).

is relatively low with 0.54, reflecting high inconsistency in human segment-level labels. As n increases, inter-annotator accuracy rises sharply. Grouping only three segments already increases human–human pairwise accuracy from 0.55 to 0.70. Metric–human agreement improves in parallel and remains close to inter-annotator agreement across all values of n. This behavior indicates that, on seen data, reducing human noise primarily strengthens agreement signals without revealing large gaps between humans and metrics.

On the unseen PharmaChem domain at $n = 1$, the gap between inter-annotator agreement and the best metric–human agreement is small (approximately 0.04), mirroring the segment-level results discussed earlier. As we increase n, inter-annotator agreement continues to rise and approaches 1.0, reflecting increasingly stable human system rankings. In contrast, metric–human agreement plateaus around 0.8. As a result, the gap between human–human and human–metric agreement widens, reaching approximately 0.20 for larger group sizes. This divergence indicates that once human label inconsistency is reduced, persistent metric weaknesses in unseen domains become more visible.

Grouping also exposes differences between meaningful metrics and trivial baselines. At $n = 1$, GEMBA performs on par with the sentinel baseline across all domains, indicating that human noise masks the benefit of access to the source sentence. When grouping at least $n = 6$ segments, this changes: GEMBA outperforms the sentinel by approximately 0.08 pairwise accuracy across all three domains. This aligns with the expectation that access to the source sentence improves translation evaluation once annotation noise is reduced. A similar pattern-no separation at $n = 1$, but a growing gap after grouping is also observed on the Bio dataset of Zouhar et al. (2024) (see Section A).

For noisy human scores, where strong metrics only marginally outperform a random baseline, we recommend selecting the minimal grouping size n (choosing $n \leq 6$ is often sufficient) that reveals a statistically meaningful separation between a random baseline and a strong metric. By choosing n like this, we remain as close as possible to segment-level evaluation while mitigating human labeling noise.

4.3 Automatic metrics are susceptible to biases on unseen domains.

In this section, we analyze the systematic biases that occur under domain shifts for various metrics.

Bias 1: QE metrics overestimate open-source systems on unseen domains. Any given QE metric should align with human judgments on the relative ranking of MT systems across domains. To assess this, we compute system rankings separately for each domain using z-scores. These ranks allow us to compare scores across heterogeneous scoring scales of different humans and metrics. For each human and metric, scores are normalized per domain, and systems are ranked by their average score across all translations.

Figure 5 shows differences between human- and metric-based rankings for three systems X-ALMA (open-source), Azure, and DeepL (closed-source), for which human judgments reveal performance gaps between open- and closed-source models in en-de. Human rankings (HumanA/B) are obtained by splitting the four annotators randomly into two groups (WMT23, Emea) or using individual annotators directly (PharmaChem).

On the seen WMT23 domain, humans and metrics largely agree. Both rank DeepL and Azure above X-ALMA (e.g., HumanA/B: DeepL +0.77, Azure +0.62 vs. X-ALMA -1.13). Metrics reflect this ordering, with SFT metrics assigning DeepL +1.35, Azure +0.51, and X-ALMA -0.71; Remedy is the only metric placing X-ALMA close to closed-

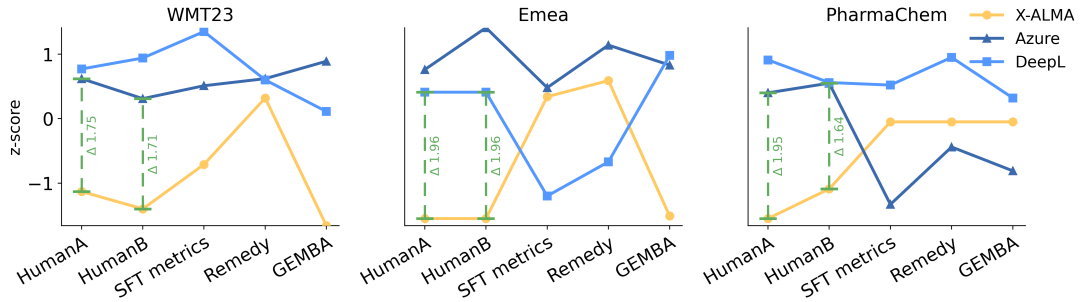


Figure 5: Visualization of the ranking of three MT systems. The x-axis lists the ranks and the y-axis shows the corresponding z-scores assigned by the ranker. Green dotted Δ markers in the HumanA/HumanB columns indicate the score gap between X-ALMA and the closed-source systems under human judgments.

source systems.

On unseen domains, this alignment breaks down. Humans consistently prefer closed-source systems, while metrics inflate scores for X-ALMA and deflate closed-source systems. On Emea, humans assign DeepL +0.41 and X-ALMA -1.55, whereas SFT metrics score X-ALMA +0.34 and DeepL -1.20, reversing the human ranking. A similar effect appears on PharmaChem: humans rank DeepL (+0.91) and Azure (+0.40) above X-ALMA (-1.55), while metrics assign X-ALMA scores near zero (e.g., -0.05 for SFT metrics) and push Azure into negative ranges (-1.33).

While the precise score shifts are not all visible in Figure 5, detailed rankings and score differences for all systems are provided in Section G.

The bias is most pronounced for supervised fine-tuned metrics, but it also appears for the LLM-as-a-judge approach on the contamination-free PharmaChem domain. Notably, these ranking errors occur despite using the same scoring procedures as those used in WMT23, and despite strong inter-annotator agreement among humans.

However, we could not observe the bias for en-zh and en-ko. This effect may be due to several factors. First, shared training data or reward alignment between QE metrics and open-source systems (e.g., X-ALMA) may be stronger for en-de, where more human-annotated data is available, potentially leading to increased system preferences. Second, systems may be more closely matched in quality for en-de, making rankings more sensitive to small scoring differences. Thus, system-specific effects (e.g., relatively weaker performance of X-ALMA in other language pairs) may reduce the visibility of the bias outside en-de.

Bias 2: QE metrics underestimate translation quality on out-of-domain data. A desirable property of a QE metric is domain invariance: translation quality should be assessed independently of the source domain. We test this by comparing average human and metric scores across domains. We use z-scores to ensure comparability across different scoring scales.

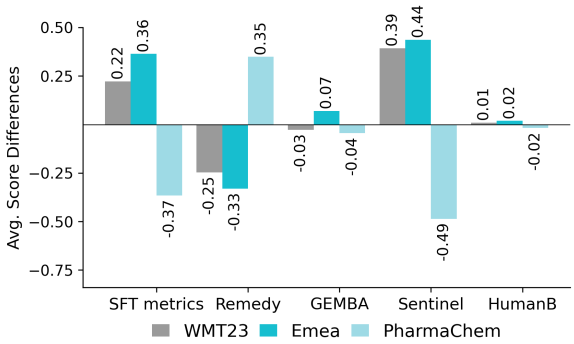


Figure 6: Average differences of normalized human and metric scores across domains. Negative values mean the metric underestimates translation quality relative to humans; positive values mean it overestimates quality.

For each domain, we compute the average z-score assigned by humans and metrics and their differences (Figure 6) over all three language pairs. Because human judgments reflect translation quality rather than domain characteristics, a domain-agnostic metric should produce average scores that closely align with human average scores across domains, resulting in differences near zero. The results for each pair of languages are in section D.

Figure 6 shows that on the unseen PharmaChem domain, SFT metrics are systematically pessimistic, assigning scores on average 0.37 points lower than humans. Their behavior closely mirrors that of the sentinel baseline, suggesting reliance on target-side fluency rather than adequacy under domain shift. Remedy exhibits inconsistent behavior across unseen domains, overestimating quality on

PharmaChem while underestimating it on Emea. In contrast, the LLM-as-a-judge approach (GEMBA) does not show systematic over- or underestimation, indicating more domain-agnostic scoring behavior.

Based on these findings, we recommend using LLM-as-a-judge when comparing translation quality across domains, as they provide more reliable estimates under domain shift.

Error masking analysis. To better understand why supervised fine-tuned (SFT) metrics exhibit the biases observed on unseen domains, we analyze token-level error masks predicted by XCOMET_QE, an explainable variant of COMET (Rei et al., 2020). Error span prediction offers a fine-grained view of metric behavior beyond scalar scores on the WMT25 Metrics shared task 2.

We compare character-level overlap between XCOMET_QE-predicted error spans and human annotations across all domains, computing precision, recall, and F1 following the WMT25 setup. For simplicity, we treat minor and major errors uniformly. Human–human agreement shows balanced precision and recall, with stable F1 scores between 0.37 and 0.47, indicating consistent annotation behavior across domains (Table 2).

In contrast, human–XCOMET_QE agreement is substantially lower (F1: 0.28–0.33), confirming that error span prediction remains challenging. Precision–recall imbalance is markedly stronger on unseen domains, most notably on PharmaChem, where precision drops to 0.23 while recall reaches 0.80. The imbalance indicates that XCOMET_QE over-predicts errors particularly under domain shift. Thus, SFT metrics may rely on source-side domain cues rather than solely judging translation quality.

Qualitative examples corroborate the pattern: on unseen-domain translations, XCOMET_QE frequently labels large portions of a sentence as containing minor errors, even when humans annotate a small number of error spans (Figure 7). Our analysis reveals a systematic tendency of SFT metrics to over-predict errors on unseen domains, providing an explanation for the pessimistic quality estimates observed in Figure 6.

Embedding Analysis. We analyze last-layer representations of XCOMET. The embeddings cluster mainly by domain rather than by human-judged quality, indicating reliance on source-side domain information. Details are in Section E.

Domain	XCOMET–human			human–human		
	P	R	F1	P	R	F1
WMT23	0.27	**0.46**	0.28	**0.42**	0.43	**0.37**
Emea	0.33	**0.51**	0.33	**0.48**	0.48	**0.43**
PharmaChem	0.23	**0.80**	0.32	**0.53**	0.53	**0.47**

Table 2: **P**recision, **R**ecall, and **F1** of character-level error mask overlap between human–human and human–XCOMET. **Bold** shows the highest value per domain.

Source:
Grease line is interrupted near the front end.
xCOMET:
Fettleitung ist nahe am Front Ende unterbrochen
Human A:
Fettleitung ist nahe am Front Ende unterbrochen
Human B:
Fettleitung ist nahe am Front Ende unterbrochen

Figure 7: Example of XCOMET on Emea marking many words as errors when humans only marked two.

5 Conclusion

This paper analyzes how automatic MT evaluation metrics behave under domain shift. Using CD-ESA, a new multi-annotator dataset covering translations from the same five MT systems on a seen news domain (WMT23) and two unseen technical domains (Emea and the proprietary PharmaChem), all annotated by the same five annotators, we study metric behavior under controlled cross-domain conditions.

We show that annotation noise can mask metric weaknesses on unseen domains, leading to overly optimistic conclusions. When we reduce noise by averaging multiple human annotations or grouping segments, marked gaps emerge between human and metric judgments on unseen domains.

On unseen domains, metrics show two systematic biases, they underestimating translation quality and over-preferencing weaker open-source systems. These effects are strongest for SFT metrics, while LLM-as-a-judge approaches show greater robustness. Error mask analyses further shows that SFT metrics rely on domain-related cues rather than translation quality.

Our findings highlight that metric robustness across domains must be interpreted relative to inter-annotator agreement under identical conditions. Controlling for human label inconsistency is essential for reliable conclusions about metrics.

Limitations

Our study focuses on two unseen technical domains (Emea and PharmaChem) and a single

high-resource language pair (English–German), reflecting both our research goal of evaluating metrics in our-of domains settings and the needs of our pharmaceutical industry partner. While extending this approach to additional domains and language pairs is an important direction for future work, the multi-annotator error span annotations required for reliable evaluation are costly at scale. Future studies could mitigate this cost by reusing existing annotations, sharing cross-domain annotation resources, or adopting aggregation strategies such as segment grouping to reduce annotation overhead.

Some of our analyses further rely on assumptions about human judgments. In particular, we assume that human annotators assess translations based on translation quality rather than domain characteristics. Under this assumption, a domain-agnostic metric should assign domain-wise average scores similar to those of humans. To our knowledge, this assumption is necessary for fair cross-domain comparison, but our conclusions may need to be revisited if future work challenges it.

Finally, reducing human annotation noise requires multiple annotations per segment, which substantially increases annotation cost. While averaging human judgments improves evaluation reliability, it is expensive in practice. Our proposed grouping strategy offers a more cost-effective alternative, but it reduces the number of independent evaluation units. To mitigate this trade-off, future evaluation campaigns could prioritize annotating shorter segments, which would allow collecting more annotations under fixed annotation budgets.

Acknowledgments

This work was supported by the Lower Saxony Ministry of Science and Culture and the VW Foundation, and by the Federal Ministry for Economic Affairs and Climate Action (BMWK) on the basis of a decision by the German Bundestag.

The authors would like to thank the German Federal Ministry of Research, Technology and Space and the German federal states (http://www.nhr-verein.de/en/our-partners) for supporting this work/project as part of the National High-Performance Computing (NHR) joint funding program.

We thank Anastasia Zhukova for supporting on the data curation and preparation phase and eschbach GmbH, specifically Christian E. Lobmüller, for funding the data annotation effort to evaluate the approach on the proprietary domain.

We thank Tianyu Yang for carefully reading the final draft of the paper and providing valuable feedback. Finally, we thank Lars Kaesberg and Jonas Becker for their support in improving the figure design.

References

Sweta Agrawal, António Farinhas, Ricardo Rei, and André F. T. Martins. 2024. Can automatic metrics assess high-quality translations? *Preprint*, arXiv:2405.18348.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4623–4637.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Anja Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, and 1 others. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in nlp. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10.

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Arun Tejasvi Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. *ArXiv*, abs/1807.02202.

Erenay Dayanık and Sebastian Padó. 2021. Disentangling document topic and author gender in multiple languages: Lessons for adversarial debiasing. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–61.

DeepL SE. 2024. Deepl translator. https://www.deepl.com/translator.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. *arXiv preprint arXiv:2305.14324*.

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. *Preprint*, arXiv:2308.07286.

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

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022a. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, and 1 others. 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.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André FT Martins. 2022b. 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.

Yvette Graham, Timothy Baldwin, and Nitika Mathur. 2015. Accurate evaluation of segment-level machine translation metrics. In *North American Chapter of the Association for Computational Linguistics*.

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.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. Gemba-mqm: Detecting translation quality error spans with gpt-4. *arXiv preprint arXiv:2310.13988*.

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. *Preprint*, arXiv:2107.10821.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. *arXiv preprint arXiv:2406.11580*.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, and 1 others. 2025. Findings of the wmt25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Microsoft. 2024. Microsoft translator documentation. https://learn.microsoft.com/azure/ai-services/translator/.

OpenAI. 2024. Gpt-4o system card. https://openai.com/research/gpt-4o-system-card.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! *arXiv preprint arXiv:2408.13831*.

Barbara Plank. 2022. The'problem'of human label variation: On ground truth in data, modeling and evaluation. *arXiv preprint arXiv:2211.02570*.

José Pombal, Nuno M Guerreiro, Ricardo Rei, and André FT Martins. 2025. Adding chocolate to mint: Mitigating metric interference in machine translation. *Transactions of the Association for Computational Linguistics*, 13:1319–1339.

Lorenzo Proietti, Stefano Perrella, and Roberto Navigli. 2025. Has machine translation evaluation achieved human parity? the human reference and the limits of progress. *arXiv preprint arXiv:2506.19571*.

Ricardo Rei, Nuno M Guerreiro, Daan Van Stigt, Marcos Treviso, Luísa Coheur, José GC de Souza, André FT Martins, and 1 others. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848.

Ricardo Rei, José Pombal, Nuno M Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, and 1 others. 2024. Tower v2: Unbabel-ist 2024 submission for the general mt shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Gabriele Sarti, Vilém Zouhar, Malvina Nissim, and Arianna Bisazza. 2025. Unsupervised word-level quality estimation for machine translation through the lens of annotators (dis)agreement. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, page 18320–18337. Association for Computational Linguistics.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2023. Scirepeval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5548–5566.

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. In *European Association for Machine Translation Conferences/Workshops*.

Emma Steigerwald, Valeria Ramírez-Castañeda, Débora Y. C. Brandt, András Báldi, J. Shapiro, Lynne Bowker, and Rebecca D. Tarvin. 2022. Overcoming language barriers in academia: Machine translation tools and a vision for a multilingual future. *Bioscience*, 72:988 – 998.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *ArXiv*, abs/1409.3215.

Shaomu Tan and Christof Monz. 2025. Remedy: Learning machine translation evaluation from human preferences with reward modeling. *arXiv preprint arXiv:2504.13630*.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. *arXiv preprint arXiv:2409.09598*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Christian Tomani, David Vilar, Markus Freitag, Colin Cherry, Subhajit Naskar, Mara Finkelstein, Xavier Garcia, and Daniel Cremers. 2024. Quality-aware translation models: Efficient generation and quality estimation in a single model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15660–15679.

J. Wahle, T. Ruas, S. M. Mohammad, N. Meuschke, and B. Gipp. 2023. Ai usage cards: Responsibly reporting ai-generated content. In *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 282–284, Los Alamitos, CA, USA. IEEE Computer Society.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2024. X-alma: Plug & play modules and adaptive rejection for quality translation at scale. *arXiv preprint arXiv:2410.03115*.

Ran Zhang, Wei Zhao, and Steffen Eger. 2025. How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10961–10988, Albuquerque, New Mexico. Association for Computational Linguistics.

Vilém Zouhar, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. *arXiv preprint arXiv:2402.18747*.

A Pairwise accuracy evaluation for the BIO-MQM dataset

Prior work by Zouhar et al. (2024) evaluates metric robustness on the BIO-MQM dataset using plain Kendall's τ for meta-evaluation. While Kendall's τ is a widely used rank-correlation measure, it does not explicitly account for ties, which are common in high-quality translation data.

More recent meta-evaluation techniques address these issues by (i) comparing systems on a

pairwise basis with explicit tie handling (Deutsch et al., 2023), and (ii) restricting comparisons to translations of the same source segment via a group-by-item evaluation strategy. Together, these methods prevent measuring spurious correlations across heterogeneous segments and consider if metrics are able to predict ties.

Applying these updated evaluation techniques to the BIO-MQM data yields a more nuanced view of metric performance and highlights several properties of the dataset that influence the resulting conclusions.

Findings from the updated BIO meta-evaluation. Figure 8 summarizes metric-human agreement under pairwise accuracy with tie calibration.

First, for the Chinese–English BIO data, performance is strongly affected by dataset characteristics. A large proportion of translations contain no annotated errors, leading to label imbalance. Under these conditions, no metric consistently outperforms a constant baseline that predicts ties for all system pairs. This suggests that apparent metric failures on this subset are driven primarily by data properties rather than by a lack of metric signal.

Second, BLEU performs unexpectedly well on the BIO domain, particularly for English–Russian, where it exceeds the constant baseline by nearly 20 accuracy points. On WMT22 news data, in contrast, BLEU performs only marginally above this baseline. This pattern indicates that BLEU's behavior is domain-dependent: it appears to benefit from the more literal and formulaic nature of biomedical text, whereas it is less effective on the more diverse and conversational content found in WMT news data. As a result, BLEU cannot be assumed to act as a domain-invariant reference metric.

Finally, we note a structural limitation of the BIO-MQM dataset as used in prior work. Only a small number of segments are translated by the same systems and annotated by at least two human annotators, which makes it difficult to compute reliable inter-annotator agreement or human-oracle upper bounds. This limits the interpretability of absolute metric performance levels on this dataset.

Grouping analysis on BIO data. To further assess whether label imbalance and annotation noise drive these effects, we apply our proposed grouping n segments strategy described in Section 4.2. As shown in Figure 9, grouping a small number of segments substantially stabilizes metric evaluation on the Chinese–English BIO data without discarding any annotations.

Under grouping, most metrics outperform a random baseline, indicating that earlier near-random performance was largely due to noise and imbalance at the segment level rather than an absence of useful metric signal. BLEU, however, continues to lag behind neural metrics under grouping, consistent with its known limitations in fine-grained quality estimation.

Overall, this analysis illustrates that conclusions about metric robustness on unseen domains are sensitive to the choice of meta-evaluation protocol and to dataset properties such as label balance and annotation density. These observations motivate the controlled evaluation design adopted in our main experiments.

B Meta-evaluation details

Segment-level. For a given segment, let C denote the number of translation pairs ranked in the same order by both m and h, and D the number ranked in opposite order. Let T_m and T_h denote pairs tied only by the metric or only by the human, and T_{mh} pairs tied by both. Segment-level accuracy is then defined as:

$$\text{acc}_{eq} = \frac{C + T_{mh}}{C + D + T_m + T_h + T_{mh}}. \quad (1)$$

Tie calibration introduces an optimal threshold on metric score differences (i.e., the maximum score difference still treated as a tie), so that metrics producing continuous scores and rarely predicting exact ties can still be fairly compared to human judgments.

System-level. Given a test set with M MT systems, SPA considers all $\binom{M}{2}$ system pairs. For each pair (i, j), we compute a human p-value p^h_{ij} and a metric p-value p^m_{ij}, representing the statistical confidence that system i is preferred over system j. These p-values are obtained via paired permutation tests over segment-level scores. SPA assigns partial credit based on how similar the metric's confidence is to the human confidence:

$$\text{SPA} = \binom{M}{2}^{-1} \sum_{i=0}^{M-1} \sum_{j=i+1}^{M-1} \left(1 - \left|p^h_{ij} - p^m_{ij}\right|\right). \quad (2)$$

Thus, SPA rewards metrics that express preference strengths similar to those of human judgments and penalizes deviations.

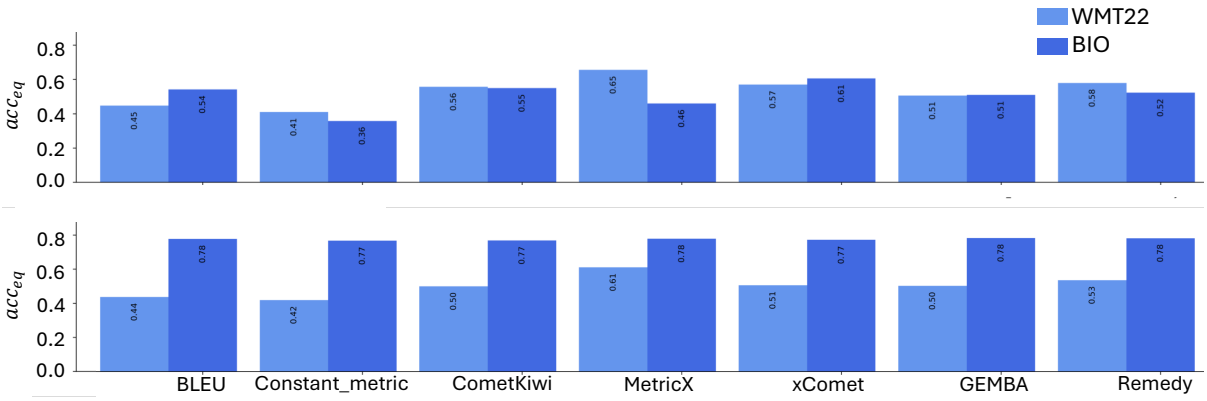


Figure 8: Pairwise metric–human accuracy on BIO-MQM data from WMT22. Except for BLEU, which requires references, all metrics are evaluated in a reference-free (QE) setting. The top panel shows results for English–Russian and the bottom panel for Chinese–English.

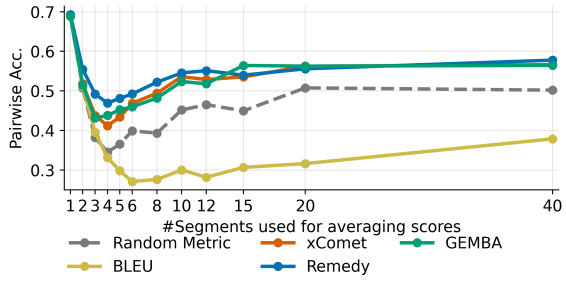


Figure 9: Pairwise accuracy on Chinese–English BIO data when grouping multiple segments into a single evaluation unit.

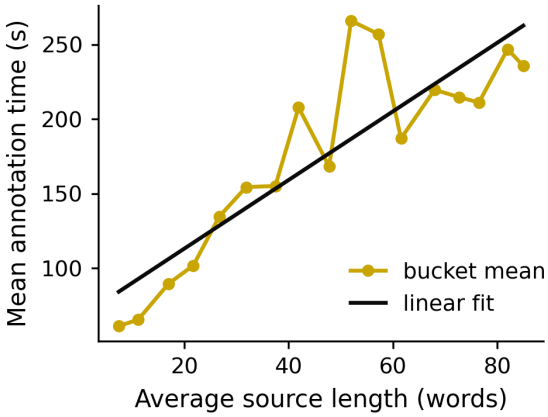


Figure 10: Human annotation time as a function of source segment length. Source sentences are bucketed by length in intervals of five words. The average segment length (x-axis) and annotation time (y-axis) is reported per bucket. The mean is computed over the 75% fastest annotations within each bucket.

C Human annotation times

Figure 10 analyzes how human annotation time scales with source segment length. We bucket source sentences by length (in word intervals of five) and report the average annotation time per bucket. To reduce the influence of outliers (e.g., breaks during annotation), we compute the mean over the 75% fastest annotations within each bucket.

The results show that annotation time increases approximately linearly with segment length. These finding supports the recommendation made in the main text: annotating shorter segments is more cost-efficient. When combined with grouping strategies that aggregate multiple short segments into a single evaluation unit, this approach enhances evaluation robustness while limiting annotation cost.

D Over pessimism bias of QE metrics on unseen domains across three langauge pairs

Following Figure 6, we compute, for each domain, the difference between average human and metric z-scores to assess whether metrics systematically over- or underestimate translation quality. Instead of averaging over all three language pairs, we report results for each language pair in Figure 11. Overall, the patterns observed in the main text are consistent across all language pairs.

On the unseen PharmaChem domain, SFT metrics consistently underestimate translation quality relative to human judgments. For all three langauge pairs the SFT metrics score the translation

on the PharmaChem between 0.3 to 0.45 below human scores. Notably, the behavior of SFT metrics closely follows that of the `sentinel_cand` baseline across all three language pairs.

Remedy again shows unstable behavior across domains. For instance, it overestimates quality on PharmaChem across all three language pairs by 0.30 to 0.48 points and underestimates translation quality by 0.28-0.39 on Emea. In contrast, GEMBA exhibits more stable behavior, with differences to human scores remaining closer to zero across domains and language pairs (typically within $\pm$0.01–0.15), indicating more domain-agnostic scoring.

These consistent trends across English–German, English–Chinese, and English–Korean suggest that the observed pessimism bias of supervised QE metrics is not language-specific but a general phenomenon under domain shift.

E Embeddings of XCOMET

To understand why SFT metrics become overly pessimistic on unseen domains, we analyze what information their internal representations encode. Specifically, we extract the last-layer embeddings of XCOMET before the scoring head and visualize translation embeddings from WMT23 and PharmaChem using t-SNE.

Ideally, these embeddings would primarily reflect translation quality. Good and bad translations should form separate clusters, and these groupings should be largely independent of the domain of the source sentence.

Figure 12 shows the resulting embedding space, where each point corresponds to a translation, and colormaps indicate the source domain. Rather than clustering by human-judged translation quality, the embeddings separate almost entirely by domain. Translations from WMT23 and PharmaChem exhibit a clear domain-dependent tendency, clustering more by domain than by human-judged quality. This pattern indicates that domain information is unsurprisingly still encoded, which could partially explain reduced robustness under domain shift.

These results indicate that XCOMET learns strong domain-specific representations. As a consequence, the metric relies heavily on source-side domain information, which weakens its ability to generalize and to judge translation quality consistently across domains.

We observe the same behavior when repeating the analysis for WMT23 and Emea (details in Section E). In all cases, domain-specific clustering persists, largely independent of translation quality.

F Data filtering

To obtain a close proxy to our private company dataset, we manually searched for domain-related multilingual and parallel open-source resources and trained a domain classifier to filter and retain only those segments that closely match the chemical and pharmaceutical domain. To train a domain classifier for identifying chemical and pharmaceutical content, we constructed a dataset with a balanced mixture of related and unrelated segments. Half of the data (50%) consists of unrelated domains such as Law, Art, Physics, and History, sampled from the EU Bookshop corpus.[6] The other half (50%) contains domain-related text. Of this portion, 50% is drawn from FoS categories Medicine, Pharma (Singh et al., 2023), and 50% consists of English instances from the PharmaChem domain, sampled randomly to ensure representation across all customers.

The final dataset contains 80K documents, split into train, dev, and test sets. We fine-tuned SciBERT (Beltagy et al., 2019) using the following hyperparameters:

```
- learning_rate=2e-5,
- batch_size=16,
- num_train_epochs=3,
- weight_decay=0.01,
- evaluation_strategy=epoch,
- save_strategy=epoch,
- load_best_model_at_end=True
```

We observed overfitting beyond 3 epochs and therefore kept the best checkpoint. The final model achieved accuracy 0.95, precision 0.97, recall 0.93, and F1 0.95.

We qualitatively inspected several multilingual datasets for domain relevance, including XQuAD (Artetxe et al., 2020) (Wikipedia QA pairs), TED Talks (Dayanık and Padó, 2021) (semi-formal spoken-language transcripts), and Emea (Tiedemann, 2012) (medical documents from the European Medicines Agency). Among these, we found that Emea is the closest in style and domain to the PharmaChem domain. After training, we then applied the classifier to the Emea corpus

[6] https://metatext.io/datasets/eubookshop

and retained only sentences with a classifier score > 0.5 to obtain a proxy dataset for the private company dataset.

G System ranking table

This appendix reports the full system rankings for english-german underlying the analyses in Section 4.3. For each domain, we show rankings assigned by individual human groups (HumanA, HumanB) and by each evaluated metric. Rankings are computed from domain-wise z-normalized scores, with higher values indicating better performance.

The tables make explicit where metrics diverge from human judgments, including cases of attenuated score differences and outright ranking reversals between open- and closed-source systems. They also provide the complete score context for systems not visualized in the main figures, enabling a detailed inspection of metric behavior across domains.

H Human annotation guidelines

Guidelines and Labeling setup were adapted from error span annotation protocol (ESA, (Kocmi et al., 2024)). Provided guidelines:

Introduction

We want to find out which translation system on the shift connector data is mostly preferred by humans, and if automatic evaluation metrics come to the same result as human evaluation. For that, translations of English source sentences were done by six different machine translation systems (Deepl, GPT4-o, Azure Translate, TowerInstruct-7B-v0.2, Tower fine-tuned on PharmaChem data, and X-ALMA), which are known to produce high-quality translations. You are asked to annotate the quality of the translated sentences.

Guidelines

Annotate the English-German translation by marking errors and providing a translation quality score:

- Choose error severity for translation:
  - Minor errors: Style, grammar, and word choice could be better or more natural.
  - Major errors: The meaning is changed significantly, and/or the part is really hard to understand.
- Missing content: If something is missing, tick either the "Minor Omission" or "Severe Omission" box. If not, tick the "No Omission" box
- Tip: Highlight the word or general area of the error—it doesn't need to be exact. Use separate highlights for different errors.
- Score the translation: After marking errors, please set an overall score based on meaning preservation and general quality:
  - 0: No meaning preserved: most information is lost.
  - 33%: Some meaning preserved: major gaps and narrative issues.
  - 66%: Most meaning preserved: minor issues with grammar or consistency.
  - 100%: Perfect: meaning and grammar align completely with the source

Tip: Assign the score such that it reflects your preferred translations. If you prefer translation of system 1 over the other two translations, the score of translation 1 should be the highest.

Additional Notes

- If you don't know an abbreviation or word, quickly Google it or ask Google Gemini.
- Do not use ChatGPT if you don't know a word, as we use ChatGPT as one of the translation systems. It probably gives you the same (but maybe wrong) translation.
- Some source sentences will appear twice. This is done on purpose for validation reasons. Proceed as always.
- The order of the translation systems is randomly shuffled. For example, the translation of system 1 could be the GPT4-o translation for one example and in the next example, system 1 could be Deepl.

Annotation interface is adapted from ESA annotation interface [7] and rebuild in Label-Studio. An image of the rebuild interface is given in Figure 13.

[7] https://github.com/AppraiseDev/Appraise

humanA	humanB	SFT metrics	GEMBA	Remedy
GPT4o (+0.94)	GPT4o (+1.00)	DeepL (+1.35)	GPT4o (+1.11)	Azure (+0.62)
DeepL (+0.77)	DeepL (+0.94)	Azure (+0.51)	Azure (+0.89)	DeepL (+0.60)
Azure (+0.62)	Azure (+0.31)	GPT4o (+0.32)	DeepL (+0.11)	GPT4o (+0.44)
ALMA (-1.13)	Tower-Instrc (-0.86)	ALMA (-0.71)	Tower-Instrc (-0.45)	ALMA (+0.32)
Tower-Instrc (-1.19)	ALMA (-1.40)	Tower-Instrc (-1.46)	ALMA (-1.66)	Tower-Instrc (-1.99)

(a) WMT23 (en–de).

humanA	humanB	SFT metrics	GEMBA	Remedy
Azure (+0.76)	Azure (+1.41)	GPT4o (+1.32)	DeepL (+0.98)	Azure (+1.14)
GPT4o (+0.74)	DeepL (+0.41)	Azure (+0.48)	Azure (+0.83)	ALMA (+0.59)
DeepL (+0.41)	GPT4o (+0.26)	ALMA (+0.34)	GPT4o (+0.56)	GPT4o (+0.56)
Tower-Instrc (-0.36)	Tower-Instrc (-0.53)	Tower-Instrc (-0.94)	Tower-Instrc (-0.86)	DeepL (-0.67)
ALMA (-1.55)	ALMA (-1.55)	DeepL (-1.20)	ALMA (-1.51)	Tower-Instrc (-1.62)

(b) Emea (en–de).

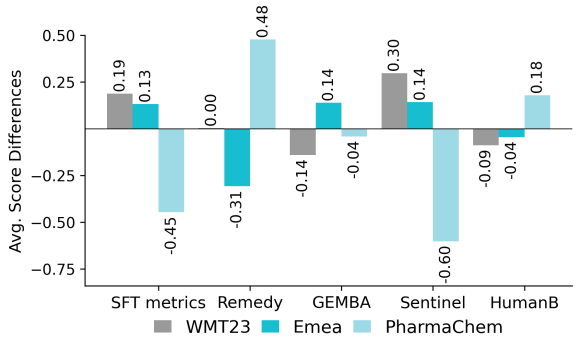
humanA	humanB	SFT metrics	GEMBA	Remedy
GPT4o (+1.00)	GPT4o (+1.26)	GPT4o (+1.48)	GPT4o (+1.70)	GPT4o (+1.15)
DeepL (+0.91)	DeepL (+0.56)	DeepL (+0.52)	DeepL (+0.32)	DeepL (+0.95)
Azure (+0.40)	Azure (+0.55)	ALMA (-0.05)	ALMA (-0.09)	ALMA (-0.05)
Tower-Instrc (-0.77)	ALMA (-1.09)	Tower-Instrc (-0.62)	Azure (-0.81)	Azure (-0.44)
ALMA (-1.55)	Tower-Instrc (-1.27)	Azure (-1.33)	Tower-Instrc (-1.16)	Tower-Instrc (-1.61)

(c) PharmaChem (en–de).

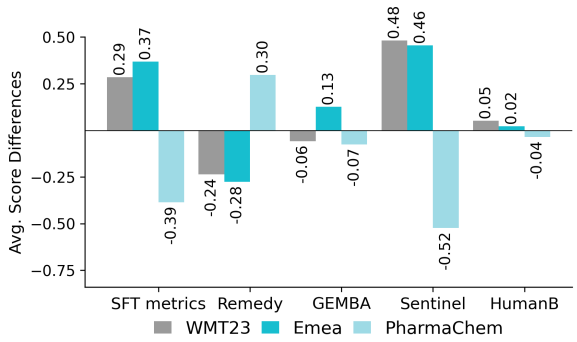
Table 3: Human and metric rankings of MT systems across three domains. Columns show rankings produced by each human group or metric; values in parentheses are average domain-wise z-normalized scores. Green highlights system pairs where humans agree but metrics do not. Orange indicates cases where metrics preserve the human-preferred ordering between open- and closed-source systems but with substantially smaller score gaps. Red marks ranking reversals relative to human judgments. For WMT23 and Emea, four annotators were split into two pairs (HumanA/HumanB); for PharmaChem, two annotators were used directly.

I AI usage card

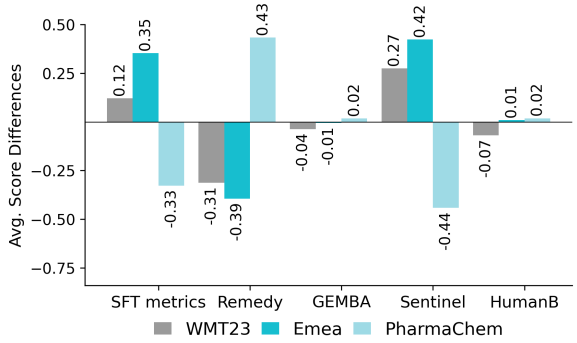
Table 4 shows how AI was used to construct this study apart from using AI models as subjects of experiments (e.g., for writing, writing code).



(a) English-German

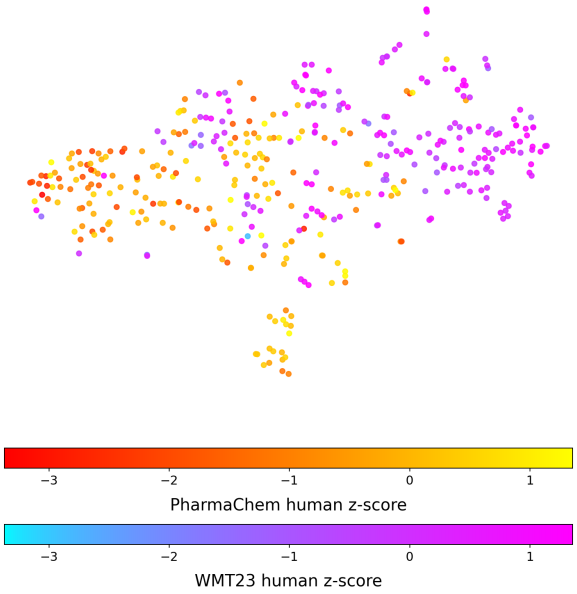


(b) English-Korean

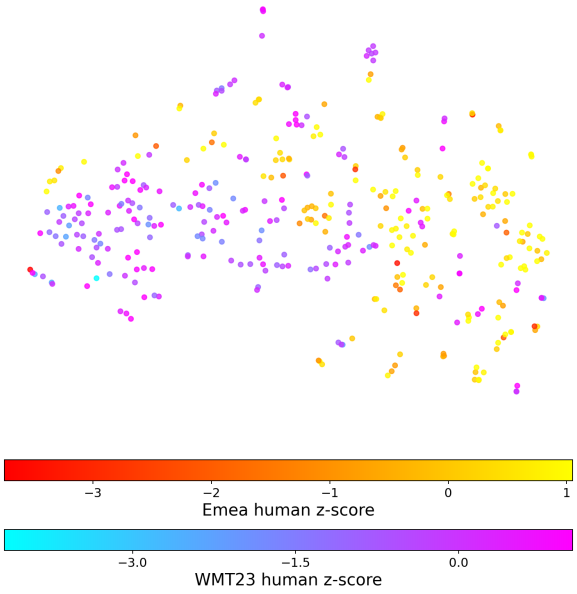


(c) English-Chinese

Figure 11: Average differences of normalized human and metric scores across domains. Negative values mean the metric underestimates translation quality relative to humans; positive values mean it overestimates quality.



(a) PharmaChem (red–yellow) vs. WMT (blue–violet).



(b) Emea (red–yellow) vs. WMT (blue–violet).

Figure 12: t-SNE visualization of last-layer translation embeddings from XCOMET (before the linear scoring head). Each point represents a translation, colored by human-assigned z-scores.

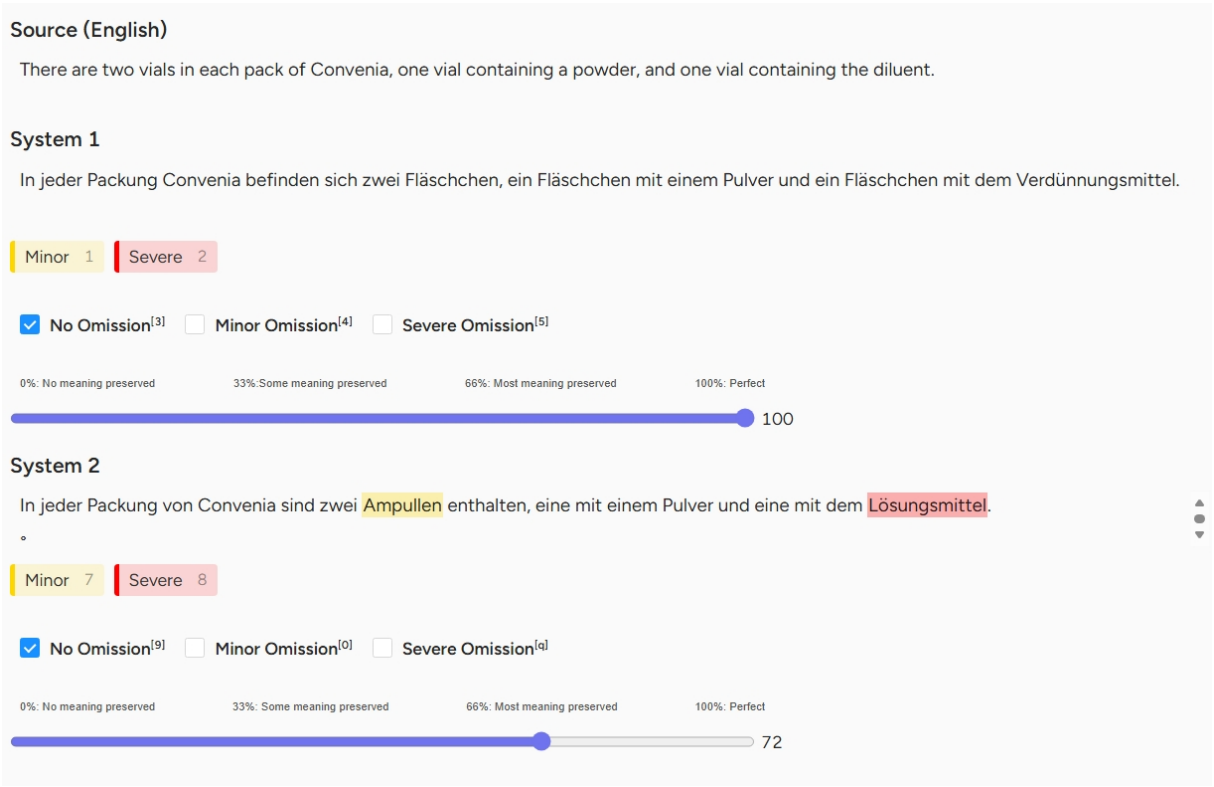


Figure 13: Example for the annotation interface in Label Studio.

AI Usage Card

PROJECT DETAILS	PROJECT NAME Who Watches the Watchmen?	DOMAIN Paper	KEY APPLICATION Coding
CONTACT(S)	NAME(S)	EMAIL(S)	AFFILIATION(S)
MODEL(S)	MODEL NAME(S) ChatGPT-5.2, GPT-5.2-Codex		
LITERATURE REVIEW	FINDING LITERATURE	FINDING EXAMPLES FROM KNOWN LITERATURE OR ADDING LITERATURE FOR EXISTING STATEMENTS	COMPARING LITERATURE
WRITING	GENERATING NEW TEXT BASED ON INSTRUCTIONS ChatGPT-5.2	ASSISTING IN IMPROVING OWN CONTENT OR PARAPHRASING RELATED WORK ChatGPT-5.2	PUTTING OTHER WORKS IN PERSPECTIVE ChatGPT-5.2
CODING	GENERATING NEW CODE BASED ON DESCRIPTIONS OR EXISTING CODE GPT-5.2-Codex	REFACTORING AND OPTIMIZING EXISTING CODE GPT-5.2-Codex	COMPARING ASPECTS OF EXISTING CODE GPT-5.2-Codex
ETHICS	WHY DID WE USE AI FOR THIS PROJECT? Efficiency and Speed	WHAT STEPS ARE WE TAKING TO MITIGATE ERRORS OF AI? We manually verified all outputs	WHAT STEPS ARE WE TAKING TO MINIMIZE THE CHANCE OF HARM OR INAPPROPRIATE USE OF AI? We carefully reviewed generated content

THE CORRESPONDING AUTHORS VERIFY AND AGREE WITH THE MODIFICATIONS OR GENERATIONS OF THEIR USED AI-GENERATED CONTENT

AI Usage Card v2.0 https://ai-cards.org (Wahle et al., 2023)